

ONLINE LEARNING AND ITS APPLICATIONS IN ELECTRICITY MARKETS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Mukadder Sevi Baltaoglu

August 2018

© 2018 Mukadder Sevi Baltaoglu

ALL RIGHTS RESERVED

ONLINE LEARNING AND ITS APPLICATIONS IN ELECTRICITY MARKETS

Mukadder Sevi Baltaoglu, Ph.D.

Cornell University 2018

Online learning is the process of learning to make accurate predictions and optimize actions sequentially in each period based on the information gained through the previous decisions and observations. In many real-world problems, the underlying model is unknown and possibly stochastic. Therefore, it is not possible to optimize actions using analytical methods. The goal of online learning is to learn from past observations as quickly as possible to minimize the loss that results from not knowing the true underlying model. Since uncertainty plays a big role in power system operations and power consumption, it makes optimizing actions a very challenging task for participants of wholesale electricity markets. This results in various interesting problems that requires an online learning approach. Motivated by two different applications in electricity markets, we study two different online learning problems.

We first study the problem of online learning and optimization of unknown Markov jump affine models which is motivated by the dynamic pricing problem of an electricity retailer. An online learning policy, referred to as Markovian simultaneous perturbations stochastic approximation (MSPSA), is proposed for two different optimization objectives: (i) the quadratic cost minimization of the regulation problem and (ii) the revenue (profit) maximization problem. It is shown that MSPSA is an order optimal learning policy in terms of regret growth rate. More specifically, the regret of MSPSA grows at the order of the square root of the learning horizon, and the regret of any policy grows no slower than that

of MSPSA. Furthermore, it is also shown that the MSPSA policy converges to the optimal control input almost surely as well as in the mean square sense. Simulation results are presented to illustrate the regret growth rate of MSPSA and to show that MSPSA can offer significant gain over the greedy certainty equivalent approach.

Motivated by virtual trading in two-settlement wholesale electricity markets, the second problem we consider is the online learning problem of optimal bidding strategy in repeated multi-commodity auctions. A polynomial-time online learning algorithm is proposed to maximize the cumulative payoff over a finite horizon by allocating the bidder's budget among his bids for K options in each period. The proposed algorithm, referred to as dynamic programming on discrete set (DPDS), achieves a regret order of $O(\sqrt{T \log T})$. By showing that the regret is lower bounded by $\Omega(\sqrt{T})$ for any strategy, we conclude that DPDS is order optimal up to a $\sqrt{\log T}$ term. Our result also implies that the expected payoff of DPDS converges, with an almost optimal convergence rate, to the expected payoff of the global optimal corresponding to the case when the underlying model is known. By using both cumulative payoff and Sharpe ratio as performance metrics, evaluations were performed based on historical data spanning ten year period of NYISO and PJM energy markets. It was shown that the proposed strategy outperforms standard benchmarks and the S&P 500 index over the same period.

BIOGRAPHICAL SKETCH

Sevi Baltaoglu graduated from Bogazici University, Istanbul, Turkey in 2013 with a Bachelor of Science degree in Electrical and Electronics Engineering. She started her graduate studies at Cornell University in 2013 and received her Master of Science degree in Electrical and Computer Engineering in 2017. Sevi Baltaoglu is currently a Ph.D. candidate at Cornell University and working under the supervision of Professor Lang Tong in the School of Electrical and Computer Engineering.

Her research interests include online machine learning, statistical learning, optimization, dynamic pricing under demand uncertainty, problems in energy markets and power systems.

For my parents, Hacer & Mustafa Baltaoglu

For my husband, Nicholas Horton

ACKNOWLEDGEMENTS

Above all, I would like to express my sincere gratitude and appreciation to my advisor, Professor Lang Tong, who patiently taught me the fundamental skills required to conduct good research. His enthusiasm, dedication, and optimism regarding research and teaching is incredible. During my graduate studies, he always gave me the encouragement and power to go further and overcome challenges at my times of desperation. I am extremely grateful to him for all of his support and the time he took to teach me how to think critically, to ask right questions, and to convey complex ideas in a simple manner.

Next, I would like to thank my committee members for their valuable inputs during my graduate studies. I am incredibly grateful to Professor Qing Zhao for her help and for her contribution to my research. She is an amazing person who always dedicates her time to discuss your research, cares a lot about every detail, and provides valuable insights whenever you need it. I must also thank to Professor Aaron Wagner not only for serving on my committee but also for teaching one of the classes that I enjoyed the most.

Besides my committee members, I am also very grateful to all other Cornell Professors who taught classes that equipped me with the necessary knowledge to conduct my research. I especially like to thank Professor Robert Kleinberg for an insightful discussion that was beneficial for part of my research.

There are so many other people who I am thankful to for cherishing my life during my Ph.D. First, I would like to thank my labmates, especially to Yuting Ji, Daniel Munoz Alvarez, Zhe Yu, Ye Guo, and Kursat Mestav, for fruitful discussions and for making our lab an enjoyable place to be. I would also like to thank other friends who I met at Cornell for making Ithaca a pleasant place to live at. I should thank separately to my first and dear friend at Cornell, Ozan

Sener, for all of his help and support at my first year at Cornell and for making the transition to Ithaca much easier for me. Also, I would like to thank my amazing friend from high school and college, Tugce Gurek, for keeping an eye on me all the time from thousands of miles away.

I am extremely grateful for everything my parents, Hacer & Mustafa Baltaoglu, and my brother, Emre Baltaoglu, did for me. I cannot express in words how important their love and support was for me to finish my Ph.D. studies. I also have to express my gratitude to my deceased aunt, Feriha Baltaoglu, who encouraged me, provided support, and gave great advice during my graduate school application process. Finally, I would like to thank my dear husband, Nicholas Horton, who I met at Cornell. I am just so grateful that he always believed in me and was there for me all the time. I really appreciate his commitment to drive from Baltimore at least twice a month for three years to visit me in Ithaca. I can't imagine what I would have done without his support.

TABLE OF CONTENTS

Biographical Sketch	iii
Dedication	iv
Acknowledgements	v
Table of Contents	vii
List of Figures	ix
1 Introduction	1
1.1 Overview of Online Learning	1
1.2 Online Learning and Optimization of Markov Jump Affine Models	5
1.2.1 Motivation: Dynamic Pricing for Demand Response	7
1.2.2 Related Work	8
1.2.3 Main Contributions	12
1.3 Online Learning of Optimal Bidding Strategy in Repeated Multi-Commodity Auctions	14
1.3.1 Motivation: Virtual Trading in Electricity Markets	15
1.3.2 Main Contributions	17
1.3.3 Related Work	18
2 Online Learning and Optimization of Markov Jump Affine Models	22
2.1 Problem Formulation	22
2.2 Online Learning for Quadratic Regulation	27
2.2.1 Optimal Solution Under Known Model and Regret	28
2.2.2 MSPSA: An Online Learning Policy	29
2.2.3 Regret Analysis for MSPSA	32
2.2.4 A Lower Bound on the Growth Rate of Regret	35
2.3 Online Learning for Revenue Maximization	38
2.3.1 Optimal Solution Under Known Model and Regret	38
2.3.2 MSPSA Policy for Revenue Maximization	39
2.3.3 MSPSA Performance and Regret Lower Bound	41
2.4 Dynamic Pricing for Demand Response	43
2.4.1 Dynamic Pricing Model	44
2.4.2 Price Responsive Demand	44
2.4.3 Online Learning for Dynamic Pricing	45
2.5 Simulation	48
2.5.1 Numerical Example for Quadratic Regulation Problem . .	50
2.5.2 Numerical Example for Revenue Maximization Problem .	51
3 Online Learning of Optimal Bidding Strategy in Repeated Multi-Commodity Auctions	54
3.1 Virtual Trading in Electricity Markets	54
3.1.1 Virtual Transactions in Two-Settlement Market System . .	54
3.1.2 A Mathematical Model of Virtual Trading	55

3.2	Problem Formulation	57
3.3	Online Learning Approach to Virtual Trading	58
3.3.1	An ERM Approach	59
3.3.2	DPDS: A Polynomial-Time Online Learning Algorithm . .	62
3.3.3	Risk-Averse Learning	64
3.4	Order Optimality of DPDS	65
3.4.1	Optimal Bidding Strategy under Known Distribution and Regret	67
3.4.2	Convergence and Regret Bound for DPDS	70
3.4.3	Lower Bound of Regret for any Bidding Policy	70
3.5	Empirical Study	71
3.5.1	Setup and Data	71
3.5.2	Benchmark Methods	72
3.5.3	Empirical Results	74
3.6	Simulation Study	80
4	Conclusions	84
4.1	Online Learning and Optimization of Markov Jump Affine Models	84
4.2	Online Learning of Optimal Bidding Strategy in Repeated Multi- Commodity Auctions	87
A	Appendix of Chapter 2	90
A.1	Proof of Lemma 1	90
A.2	Proof of Theorem 2	92
A.3	Proof of Theorem 3	93
A.4	Proof of Lemma 2	98
A.5	Proof of Theorem 7	99
B	Appendix of Chapter 3	101
B.1	Proof of Theorem 8	101
B.2	Proof of Theorem 9	105
	Bibliography	109

LIST OF FIGURES

2.1	Online learning of Markov jump affine model	25
2.2	MSPSA algorithm pseudocode for quadratic regulation problem	30
2.3	MSPSA algorithm pseudocode for revenue maximization problem	40
2.4	Average performance of MSPSA and Greedy LSE for quadratic regulation example.	49
2.5	Average performance of MSPSA and Greedy ILS for revenue maximization example.	52
3.1	Example of a piece-wise constant average payoff function of option k	61
3.2	Example of the discretization of the decision space for option k when $t = 4$	63
3.3	DPDS algorithm pseudo-code.	66
3.4	Cumulative profit trajectory from 2012 to 2016 in NYISO for $B = \$250,000$ after an initial training with 2011 data.	74
3.5	Annual performance in NYISO for $B = \$250,000$ (For each year, an initial training with previous year's data was performed.) . .	75
3.6	Annual performance in PJM for $B = \$250,000$ (For each year, an initial training with previous year's data was performed.)	76
3.7	2016 Performance in NYISO under different budget levels after an initial training with 2014 and 2015 data	77
3.8	Regret with respect to \sqrt{t} when $B = 13.845$	80
3.9	Regret with respect to \sqrt{t} when $B = 17.018$	81
3.10	Regret with respect to \sqrt{t} when $B = 20.870$	81
3.11	Regret with respect to \sqrt{t} when $B = 25.828$	82

CHAPTER 1

INTRODUCTION

Online learning has been studied in various fields such as statistics, computer science, operations research, and economics in different contexts such as clinical trials, online advertisement auctions, dynamic pricing, and repeated games. In this chapter, we first provide a general overview of online learning. Then, we will talk about the two different online learning problems studied in this thesis. For each problem, we present its motivating application in electricity markets, literature review of related work, and our contributions.

1.1 Overview of Online Learning

Online learning is the process of learning to make accurate predictions and/or optimize actions sequentially in each period based on the previous decisions and observations. The complete knowledge of the true underlying model is not available to the learner as in all machine learning problems. The goal is to adapt and improve predictions about the true underlying model each period based on the observations and additional information that becomes available that period.

In the simplest form of the online learning problem, there are finite number of actions to choose from for the learner. Depending on how the rewards that are associated with the available actions are generated, an online learning problem belongs either to the stochastic setting or to the nonstochastic (adversarial) setting. In the stochastic setting, the reward of each action is drawn from a fixed but unknown distribution at each period; whereas, in the nonstochastic setting, the reward associated with each action is assumed to be generated by

an opponent (adversary) at each period.

Online learning problems can be divided into three categories based on the information that becomes available to the learner each period. If the learner can observe the reward (or loss) associated with every action in each period regardless of the learner's own action, then the problem is called experts (or full-feedback) problem [24, 25, 34, 46]. Otherwise, it is called partial-feedback problem [25, see Chapter 6]. If the learner's observation is restricted only to the reward of the action taken by the learner at each period, then the problem is referred to as a multi-armed bandit problem [6, 55] which is a special case of partial-feedback (partial-information) setting. (See [23] for a comprehensive survey on multi-armed bandit problems.)

The fundamental difficulty in a multi-armed bandit problem (and also in most partial-feedback settings) is the exploration and exploitation tradeoff. Since the learner's goal is to optimize the total payoff over a finite or infinite horizon, the myopic approach of trying to maximize only the current period's reward fails to be optimal. The learner needs to establish a balance between exploiting what is learned so far by choosing the best action based on previous observations and exploring the outcome of other actions that are not explored enough to see if there is a better action that will lead to higher rewards in the future periods.

If the action (decision) space is a continuous set rather than a finite set, then some assumptions on the relationship between the rewards of all actions are essential. Otherwise, the learning problem becomes an impossible task due to the infeasibility of exploring the reward of each action (or even calculating the reward of each action in the case of experts problem). In general, this relationship

is established by expressing the reward of an action x_t at period t as a function of that action, *i.e.*, $r_t(x_t)$.

In the literature, the setting of continuous action space with concave (expected) reward function is studied extensively, and efficient convex optimization methods are adapted to obtain online learning algorithms with good performance guarantees such as stochastic (online) gradient descent [73, 84] and stochastic approximation methods [43, 70, 74]. In the case of non-concave but smooth reward functions, the decision space is explored by using various discretization approaches [1, 45, 47] due to the necessity of a *global search* strategy. However, for the experts problem, this approach may lead to computationally expensive algorithms that are impractical in practice in the case of high dimensional action spaces.

An online learning algorithm (policy) μ is defined as a sequence of decision rules $\mu_1, \mu_2, \mu_3, \dots$ such that at period t , μ_t maps the information history I_t containing all past observations, decisions, and any other information available to the learner at period t to the next action x_{t+1} that needs to be taken by the learner.

The performance of an online learning algorithm is generally measured by its *regret*. For the stochastic setting, regret of a policy μ is defined as the difference between the cumulative expected reward of the policy μ and that of the optimal solution under known model*. For example, for online learning problems with finite optimization horizon T , the regret is expressed as

$$\mathcal{R}_T^\mu = \sup_x \mathbb{E} \left(\sum_{t=1}^T r_t(x) - \sum_{t=1}^T r_t(x_t^\mu) \right),$$

where x_t^μ denotes the action chosen by policy μ for period t . The expectation

*The regret defined here is also referred to as pseudo-regret in some literature to reserve the term "regret" for a different definition, *e.g.*, [23].

is taken with respect to the randomness in the reward and also in the policy μ . For the adversarial setting, the same regret expression is used. However, in that case, the expectation is taken with respect to the randomness in the policy μ and in the opponent's strategy.

In the stochastic setting, since the reward distribution for a given action is fixed over time, the incremental regret is non-negative at each period. Therefore, the regret grows monotonically with T for any policy. For a simple fixed strategy that chooses the same action at each period, the regret grows linear in T . Hence, the goal is to obtain online learning policies that has a sublinear and optimally the slowest possible regret growth rate in terms of T .

To show that the regret growth rate is optimal in terms of T for an online learning policy μ , one needs to obtain an upper bound for regret of policy μ and a lower bound for regret that holds for any online learning policy and matches the regret growth rate of the upper bound in terms of T . Since the regret in the stochastic setting is a function of the model and/or distribution parameters, two different approaches are pursued in the literature to obtain bounds for regret: (i) model-dependent (distribution-dependent) approach or (ii) model-independent (distribution-free) approach. In the model-dependent approach, the upper and lower bounds for regret is derived as a function of these model and/or distribution parameters of the problem instance. In the model-independent approach, the worst case analysis, which evaluates any given policy in terms of how well it performs under the worst problem instance for that policy, are used to obtain the upper and lower bounds for regret.

Motivated by two different applications in electricity markets, we study two different online learning problems. The first one is the online learning and op-

timization problem of Markov jump affine models, which falls in the category of multi-armed bandit problems with continuous action space and convex loss (or concave reward) function. The second problem studied is the online learning of optimal bidding strategy in repeated multi-commodity auctions, which falls in the category of experts problems with high dimensional continuous action space and non-concave reward. In the former one, the difficulty lies in the exploration and exploitation dilemma. In the latter one, the difficulty is to construct an optimal online learning algorithm that is computationally efficient in high dimensions. For both problems, we consider the stochastic setting with finite optimization horizon and construct online learning algorithms with optimal regret growth rates. All of the regret bounds are derived via the model-independent approach.

This thesis incorporates materials from five different papers of the author [10–14], which are coauthored with Lang Tong and Qing Zhao. The part of this thesis that is related to the online learning and optimization of Markov jump affine models is based on the contents of [11, 12, 14], and the part of this thesis that is related to the online learning of optimal bidding strategy in repeated multi-commodity auctions is based on the contents of [10, 13].

1.2 Online Learning and Optimization of Markov Jump Affine Models

We consider the problem of online learning and optimization of affine memoryless models with unknown parameters that follow a Markov jump process. By online learning and optimization we mean that the control input of the un-

known model is chosen sequentially to minimize the expected total cost or to maximize the expected cumulative reward procured over a time horizon T . In this context, the online learning problem is one of exploration and exploitation; the need of exploring the space of unknown parameters must be balanced by the need of exploiting the knowledge acquired through learning.

The performance of an online learning policy is measured by the commonly used performance metric, regret, which is defined by the difference between the cumulative cost/reward of an online learning policy and that of a decision maker who knows the model completely and sets the input optimally. As mentioned previously, the regret grows monotonically with the time horizon T , and the rate of growth measures the efficiency of online learning policies.

The online learning problem considered here is particularly relevant in dynamic pricing problems when the consumers' demand is unknown and possibly varying stochastically [16, 39, 42, 58]. The goal of dynamic pricing is to set the price sequentially, using the observations from the previous sales, to match a certain contracted demand. Besides applications in dynamic pricing, results are also relevant to the learning and control problem of Markov jump linear systems with unknown parameters [27, 30, 59].

We study the online learning and optimization problem of Markov jump affine models under two different objectives: (i) target matching with a quadratic cost and (ii) revenue (profit) maximization. Our goal is to establish fundamental limits on the rate of regret growth for Markov jump affine models and develop an online learning policy that achieves the lowest possible regret growth.

1.2.1 Motivation: Dynamic Pricing for Demand Response

Demand response is a key component of an efficient energy market due to its economical and environmental benefits both for system operators and consumers [65]. By means of demand response, such as dynamic pricing or other incentive based methods, system operator can directly control or indirectly influence consumer demands to reduce peak demand and minimize the risk of supply outages.

We consider the use of dynamic pricing in a distribution system by a retailer such as an energy aggregator or a local utility in the two-settlement wholesale market framework [39]. In the day-ahead market, the independent system operator receives bids from generators and retailers and determines the optimal day-ahead dispatch of the next day by solving an optimal power flow problem. In the real-time market, independent system operator adjusts the day-ahead dispatch according to the real-time operating conditions, and the real-time wholesale price compensates deviations from the day-ahead schedule.

A retailer is exposed to risks from uncertainties in the wholesale market and demands of price elastic consumers. On the one hand, the retailer has to commit to purchase a certain amount of energy at the day-ahead price, and the real-time fluctuation of the wholesale price represents an unpredictable operating cost. On the other hand, the consumer it serves adjusts its consumption based on the set price by the retailer, and how a consumer responds to set prices is influenced by both the retail price and the exogenous randomness at the time of consumption such as unexpected changes of weather conditions. Without knowing how a consumer responds to prices, setting the retail price optimally seems futile.

The Markov jump affine model studied here maps to the demand curve of price-responsive electricity consumers with control input of the model corresponding to the retail price of electricity. The affine model for demand arises from the optimization problem of the consumer based on thermal load dynamics [39], and the Markov jump process models the unexpected changes in demand associated with exogenous factors.

1.2.2 Related Work

Without Markov jump as part of the model, *i.e.*, when there is a single state, the problem considered here is the classical problem of control in experiment design studied by Anderson and Taylor [4]. Anderson and Taylor proposed a certainty equivalence rule where the input is determined by using the maximum likelihood estimates of system parameters as if they were the true parameters. Despite its intuitive appeal, the Anderson-Taylor rule was shown to be suboptimal for the quadratic regulation problem by Lai and Robbins in [54] and also for the revenue maximization problem by den Boer and Zwart in [29]. In fact, there is a non-zero probability that the Anderson-Taylor rule produces an input which converges to a suboptimal value for both cases; therefore, this rule results in *incomplete learning* and a linear growth of regret.

For the scalar model in which the quadratic cost of the regulation problem is to be minimized, Lai and Robbins [52] showed that a Robbins-Monro stochastic approximation approach achieves the optimal regret order of $\Theta(\log T)$. Later, Lai and Wei [53] showed that this regret order is also achievable for a more general linear dynamic system by an adaptive regulator that uses least square esti-

mates of a reparametrized model and ensures convergence via occasional uses of white-noise probing inputs. The result was further generalized by Lai [51] to multivariate linear dynamic systems with a square invertible system matrix. The special case of the problem considered by Lai [51] is studied in [39] in the context of retail pricing of electricity under unknown demand. The authors of [39] also proposed a Robbins-Monro type of technique to achieve the optimal regret rate of $\Theta(\log T)$. Our result can be viewed as a generalization of this line of work to allow both time-varying linear models and time-invariant models with a non-invertible system matrix.

The problem studied here also falls into the category of continuum-armed bandit problem where the control input is chosen from a subset of \mathbb{R}^n with the goal of minimizing expected cost (or maximizing expected reward) that is an unknown continuous function of the input. This problem was introduced by Agrawal [1] who studied the scalar problem and proposed a policy that combines certainty equivalence control with Kernel estimator-based learning. Agrawal showed that this policy has a regret growth rate of $O(T^{3/4})$ for a uniformly Lipschitz expected cost function. Later, Kleinberg [45] proved that the optimal growth rate of regret for this problem cannot be smaller than $\Omega(T^{2/3})$ and proposed a policy that achieves $O(T^{2/3}(\log(T))^{1/3})$. Kleinberg [45] also considered the multivariate problem, *i.e.*, $n > 1$, and showed that an adaptation of Zinkevich's greedy projection algorithm achieves the regret growth rate of $O(T^{3/4})$ if the cost function is smooth and convex on a closed bounded convex input set.

Within the continuum-armed bandit formulation, the work of Cope [26] is particularly relevant because of its use of stochastic approximation to achieve

the order-optimal regret growth of $\Omega(\sqrt{T})$ for a different class of cost functions. Cope's results (both the regret lower bound and the Kiefer-Wolfowitz technique), unfortunately, cannot be applied here because of the time-varying Markov jump affine models treated here. Also relevant is the work of Rusmevichientong and Tsitsiklis [71] on the so-called linearly parameterized bandit problem where the objective is to minimize a linear cost with input selected from the unit sphere. A learning policy developed in [71] is shown to achieve the lower bound of $\Omega(\sqrt{T})$ using decoupled exploration and exploitation phases. Even though our model is similar to the one in [71] in terms of the observed output being a linear function of the input, in our problem, the unknown model parameters follow a Markov jump process and the specific cost functions studied are quadratic; thus the problem objective is different.

There is a considerable amount of work on dynamic pricing problem with the objective of revenue maximization under a demand model uncertainty in different areas such as operations research, statistics, mathematics, and computer science. In [44], a multi-armed bandit approach with a regret growth rate of $O(\sqrt{T \log T})$ was proposed for a nonparametric formulation of the problem. See also [22] where the same problem under a general parametric demand model is considered and a modified version of myopic maximum likelihood based policy is shown to achieve the regret order of $O(\sqrt{T})$, and [29] where a similar result is obtained for a class of parametric demand models. In both [44] and [22], authors proved that the lower bound for regret growth rate is $\Omega(\sqrt{T})$.

Besides more general classes of demand models, affine model similar to the one in our work has been also studied extensively; *e.g.*, [16, 42, 58]. In both [16] and [58], it is shown that approximate dynamic programming solutions may

outperform greedy method numerically. A special case of our formulation of revenue maximization problem without any Markov jump characteristics (with time-invariant model parameters) is previously investigated by Keskin and Zeevi [42]. Keskin and Zeevi proposed a semi-myopic policy that uses orthogonal pricing idea to explore and learn the system. They showed that the lowest possible regret order is $\Omega(\sqrt{T})$ for any policy, and their semi-myopic policy achieves this bound up to a logarithmic factor; *i.e.*, $O(\sqrt{T} \log T)$.

Even though the system model is assumed to be time-invariant in most of the literature, there is a considerable amount of work especially in dynamic pricing that deals with time-varying demand models due to unpredictable environmental factors affecting demand; *e.g.*, see [9] for a demand model that evolves according to a discrete state space Markov chain in a revenue management with finite inventory problem, and [15] for a dynamic programming formulation of a profit maximization problem with an unknown demand parameter following an autoregressive process. See also [41] for a revenue maximization problem with an affine demand model where the model parameters are time-varying, yet the cumulative change in the model parameters over the time horizon T is bounded. Since Keskin and Zeevi [41] measure the regret of a policy by the difference between the cumulative cost of the policy and that of a clairvoyant who knows all the future temporal changes exactly and chooses the optimal action, their characterization of regret is too pessimistic for the Markov jump model considered here.

Some other examples of related work on online learning with time-varying models apart from dynamic pricing are [17] and [82]. In [17], Besbes, Gur, and Zeevi studied the online learning problem of more general time-varying cost

functions where the cumulative temporal changes is restricted to a budget similar to [41]. However, their characterization of regret is also similar to [41] and thus, incomparable with the one in our work. Yin, Ion, and Krishnamurthy [82] also considered the problem of estimating a randomly evolving optimum of a cost function which follows a Markov jump process. Their analysis deals with the convergence of the estimate obtained via stochastic approximation to the limit (stationary) solution, whereas here, we are concerned about estimating the optimum of the cost function at each time instant given the previous state of the Markov chain. Moreover, different than our work, their analysis relies on the availability of the noisy observations of the cost function gradient and they do not characterize regret.

1.2.3 Main Contributions

Our main contribution is the generalization of online learning of time-invariant affine models to that of Markov jump affine models. It is important to note that existing online learning algorithms that are used for time-invariant affine models (e.g., [26,51,52]) are no longer applicable for Markov jump affine models because the optimal solution becomes a function of the observed state of the Markov chain, and direct implementations of existing algorithms do not take into account this observation.

For the generalized model, we propose an online learning policy, referred to as Markovian simultaneous perturbations stochastic approximation (MSPSA). By introducing the idea of state tracking, MSPSA extends Spall’s stochastic approximation method [74] to the optimization problem of an objective func-

tion that evolves according to a Markov jump process. We show that MSPSA achieves the optimal regret order of $\Theta(\sqrt{T})$ for two different objective functions studied for the affine model: (i) quadratic regulation and (ii) the revenue maximization. Furthermore, we also show that the control input of MSPSA policy converges to the optimal solution both with probability one and in mean square as $T \rightarrow \infty$. Therefore, the proposed policy eventually learns the optimal solution.

A key implication of our results is that, in comparing with Lai's result on the learning problem of a time-invariant affine model with the quadratic regulation objective [51], modulating a linear model by a Markov jump process introduces substantial learning complexity; hence, the regret order increases from $\Theta(\log T)$ to $\Theta(\sqrt{T})$. As a special case, we also show that, even in the absence of Markov jump, when the system matrix is full column rank but not invertible, the best regret order is also $\Theta(\sqrt{T})$. It is worth noting that adding just one row to a square and invertible matrix can change the worst case regret from $\Theta(\log T)$ to $\Theta(\sqrt{T})$.

The results presented here are obtained using several techniques developed in different contexts. The MSPSA policy is a generalization of Spall's stochastic approximation method to the optimization problem of an objective function following a Markov jump process. To show the optimality of MSPSA, we use the van Trees inequality [35] to lower bound the estimation error for any policy, which is a technique used in the literature previously [19, 42]. Lastly, a result on the convergence of non-negative almost supermartingales [69] is used to obtain the convergence result for MSPSA policy.

1.3 Online Learning of Optimal Bidding Strategy in Repeated Multi-Commodity Auctions

Motivated by virtual trading in the wholesale electricity markets, we consider the problem of optimal bidding in a multi-commodity uniform-price auction [63], which promotes the law of one price for identical goods. Uniform-price auction is widely used in practice. Besides virtual trading in electricity market, which is discussed in detail in the following section, examples include spectrum auction, the auction of treasury notes, the auction of emission permits (UK).

A mathematical abstraction of multi-commodity uniform-price auction is as follows. A bidder has K options (goods) to bid on at an auction. With the objective to maximize his T-period expected profit, at each period, the bidder determines how much to bid for each option subject to a budget constraint.

In the bidding period t , if a bid for option k is greater than or equal to its *auction clearing price* at that period, then the bid is cleared, and the bidder pays its auction clearing price. His revenue resulting from the cleared bid will be the option's *spot price* (utility) at period t . Hence, the payoff obtained from the cleared bid for option k is determined by the difference between auction clearing and spot prices of option k . If the bid for option k is smaller than its auction clearing price, then the bid is not cleared at that period and the resulting payoff will be zero.

We assume that the auction clearing and spot prices of all K options are drawn from an unknown joint distribution. At the end of each period, the bidder observes the auction clearing and spot prices of all options. Therefore, be-

fore choosing the bid of period t , all the information the bidder has is a vector containing previously observed auction clearing and spot prices. At each period, a bidding policy determines the next period's bid solely based on this observation history. The performance of any bidding policy is measured by its regret defined by the difference between the total expected payoff of that bidding policy and that of the optimal bidding strategy under known (auction clearing and spot) price distribution.

1.3.1 Motivation: Virtual Trading in Electricity Markets

The wholesale electricity market in the United States consists of a day-ahead and a real-time markets. Market participants submit their bids to buy (and offers to sell) electricity to the day-ahead market approximately one day ahead of time. The bids and offers cleared in the day-ahead market are financially binding. The market clearing process sets the day-ahead prices for each hour of the day and at each location of the network.

In the real-time market, the load (thus the generation) may not match to the cleared amount in the day-ahead market, and the real-time prices of electricity may also be different from their day-ahead counterparts due to a variety of reasons, including the unexpected levels of demand and supply, unplanned outages, unpredictable weather conditions [56], and possibilities of market participants exercising market power [68].

Price discrepancies between the day-ahead and real-time markets represent a form of market inefficiency. To promote *price convergence* between the two markets, in early 2000s, virtual trading was introduced in the U.S. electricity mar-

kets. Virtual trading is a financial mechanism that allows market participants and external financial entities to arbitrage on the differences between day-ahead (auction clearing) and real-time (spot) prices. Currently, cleared virtual transactions represent a significant fraction of total energy trade. In 2013, the cleared virtual transactions in the five major electricity markets was 13% * of the total load [66].

Empirical and theoretical studies have shown that increased competition due to virtual trading results in price convergence, thus improving market efficiency [36–38, 56, 68, 72, 76, 81]. Particularly, it has been argued in [76] that a virtual trader makes profit if and only if his participation drives the day-ahead and real-time price difference toward zero. Hence, to reach the socially optimal dispatch level, it is important that the virtual traders bid optimally. However, the day-ahead and real-time wholesale prices are random due to uncertainties in demand, supply, and operation conditions. Therefore, in order to learn the optimal trading strategy, a virtual trader needs to update his belief using all the new information, which allows him to adapt his bid accordingly each day.

In an electricity market, there are potentially thousands of trading options. Due to system congestion and losses, electricity prices vary in time and across locations. The goal of this work is to develop an online learning approach to virtual trading where the trader, who is constrained by a certain budget, aims to determine profitable trading options and distribute his budget among them. By online learning we mean that bids are constructed sequentially and adaptively based on the new information available. In particular, we consider the objective of maximizing the expected total payoff as well as one that involves a mean-variance type of risk measure.

*This number goes up to 38% with the inclusion of up-to-congestion transactions of PJM.

1.3.2 Main Contributions

The main contribution of this work is a polynomial-time online learning algorithm with order-optimal regret rate to the algorithmic bidding problem under budget constraints in repeated multi-commodity auctions. This result is also generalized to an objective based on a form of mean-variance risk measure. The proposed approach falls in the category of empirical risk minimization also referred to as the follow the leader approach. The main challenge here is that optimizing the payoff (risk) amounts to solving a multiple-choice knapsack problem that is known to be NP hard [40]. Referred to as dynamic programming on discrete set (DPDS), the proposed approach is inspired by a pseudo-polynomial dynamic programming approach to 0-1 Knapsack problems. DPDS allocates the limited budget of the bidder (virtual trader) among K options in polynomial time both in terms of the number of options K and in terms of the time horizon T .

Note that obtaining the optimal bidding strategy with known price distribution is itself nontrivial due to the non-convexity of the problem. Our result provides an algorithmic bidding strategy that converges to the global optimal bidding strategy. We show that the expected payoff of DPDS converges to that of the optimal strategy under known distribution by a rate no slower than $\sqrt{\log t/t}$ which results in a regret upper bound of $O(\sqrt{T \log T})$. By showing that, for any bidding strategy, the regret is lower bounded by $\Omega(\sqrt{T})$, we prove that DPDS is order optimal up to a $\sqrt{\log T}$ term.

A significant part of this work is to evaluate the performance of the proposed strategy empirically using historical data spanning the time period between 2006 and 2016 of NYISO and PJM energy markets. Extensive empirical analysis

show that the proposed strategy consistently outperforms benchmark heuristic methods, derived from other machine learning approaches, and achieves significant profit. It is worth noting that our empirical results also show that PJM and NYISO wholesale electricity markets are both profitable although PJM market presents better opportunities to virtual traders compared to NYISO.

1.3.3 Related Work

Relevant literature falls into two categories. The one that is more directly related to our work focuses on developing online learning algorithms to similar problems in the machine learning literature. The second one focuses on understanding the effects of virtual trading on the two-settlement electricity market.

The problem formulated here can be viewed in multiple machine learning perspectives. We highlight below several relevant existing approaches. Since the bidder can calculate the reward that could have been obtained by selecting any given bid value regardless of its own decision, our problem falls into the category of full-feedback version of multi-armed bandit problem, referred to as experts problem, where the reward of all arms (actions) are observable at the end of each period regardless of the chosen arm. For the case of finite number of arms, Kleinberg et al. [46] showed that, for stochastic setting, constant regret is achievable by choosing the arm with the highest average reward at each period. A special case of the adversarial setting was studied by Cesa-Bianchi et al. [24] who provided matching upper and lower bounds in the order of $\Theta(\sqrt{T})$. Later, Freund and Shapire [34] and Auer et al. [7] showed that the Hedge algorithm, a variation of weighted majority algorithm [57], achieves the matching bound for

the general setting. These results, however, do not apply to experts problems with continuous action spaces.

The stochastic experts problem where the set of arms is an uncountable compact metric space (\mathcal{X}, d) rather than finite was studied by Kleinberg and Slivkins [47] (see [48] for an extended version). Since there are uncountable number of arms, it is assumed that, in each period, a payoff function drawn from an i.i.d. distribution is observed rather than the individual payoff of each arm. Under the assumption of Lipschitz expected payoff function, they showed that the instance-specific regret of any algorithm is lower bounded by $\Omega(\sqrt{T})$. They also showed that their algorithm—NaiveExperts—achieves a regret upper bound of $O(T^\gamma)$ for any $\gamma > (b+1)/(b+2)$ where b is the isometry invariant of the metric space. Our problem is a special case of the setting studied by Kleinberg and Slivkins [47]. Unfortunately, the computational complexity of NaiveExperts grows exponentially with the dimension (number of options in our case). Therefore, it becomes computationally intractable in practice. Also, the regret lower bound in [47] doesn't provide a bound for our problem with a specific payoff. Krichene et al. [50] studied the adversarial setting and proposed an extension of the Hedge algorithm, which achieves $O(\sqrt{T \log T})$ regret under the assumption of Lipschitz payoff functions. For our problem, it is reasonable to assume that the expected payoff function is Lipschitz; yet it is clear that, at each period, the payoff realization is a step function which is not Lipschitz. Hence, Lipschitz assumption of [50] doesn't hold in our setting.

Stochastic gradient descent methods, which have low computational complexity, have been extensively studied in the literature of continuum-armed bandit [26,33,49]. However, either the concavity or the unimodality of the expected

payoff function is required for regret guarantees of these methods to hold. This may not be the case in our problem depending on the underlying distribution of prices.

A relevant work that takes an online learning perspective for the problem of a bidder engaging in repeated auctions is Weed et al. [80]. They are motivated by online advertising auctions and studied the partial information setting* of the same problem as ours, but without a budget constraint. Under the margin condition, *i.e.*, the probability of auction price occurring in close proximity of mean utility (spot price) is bounded, they showed that their algorithm, inspired by the UCB1 algorithm [6], achieves regret that ranges from $O(\log T)$ to $O(\sqrt{T \log T})$ depending on how tight the margin condition is. They also provided matching lower bounds up to a logarithmic factor. However, the analysis in [80] on the lower bound of regret does not hold for the full information setting we study here. Furthermore, their algorithm cannot be used here due to the budget constraint. Also, we do not rely on the margin condition.

Some other examples of literature on online learning in repeated auctions studied the problem of an advertiser who wants to maximize the number of clicks with a budget constraint [2, 77], or that of a seller who tries to learn the valuation of its buyer in a posted price auction [3, 64]. The settings considered in those problems are considerably different from that studied here in the implementation of budget constraints [2, 77], and in the strategic behavior of the bidder [3, 64].

Indirectly related to this work are works that analyze the impact of virtual transactions on the overall market efficiency. Theoretical analysis on the im-

*In their setting, the bidder observes market prices only if his bid is accepted at that period.

pact of virtual trading was conducted in [76] and [61] from a game theoretic perspective. Under a single trading location model, these papers analyzed the Nash equilibrium behavior of virtual traders who have their fixed individual beliefs about the market. Tang et al. [76] showed that, under Nash equilibrium, if the belief of virtual traders is correct on average, the price difference between day-ahead and real-time converges to zero as the number of virtual traders increases. Mather, Bitar, and Poolla [61] presented a simple learning strategy that guarantee convergence to the Nash equilibrium. However, convergence to Nash equilibrium doesn't guarantee price convergence. Different from the problem of learning the Nash equilibrium in a game theoretic environment with fixed beliefs [61], we study the online learning problem of a virtual trader who updates his belief each day using new observations of day-ahead and real-time prices in order to converge to the optimal trading strategy. Furthermore, we require that not only the bidding policy converges to the optimal policy but also the convergence rate is order-optimal.

Among empirical studies, [56] and [38] are the most relevant to our work. Both evaluate market efficiency before and after virtual trading by searching for profitable trading strategies. More specifically, in [56], a chance constraint portfolio selection problem was solved by estimating the distribution of day-ahead and real-time price difference, modeled as Gaussian mixture hidden Markov model, to determine the trading strategy, whereas, in [38], hypothesis testing is used to determine the existence of a profitable trading strategy at each location. Empirical analysis using CAISO data shows that virtual trading increases efficiency but the market is still inefficient. Some of the other interesting empirical studies on the impact of virtual trading are [18, 21, 72], and [36].

CHAPTER 2

ONLINE LEARNING AND OPTIMIZATION OF MARKOV JUMP AFFINE MODELS

We study the online learning and optimization of Markov jump affine models for two different objective functions in this chapter. We start with formulating the problem. We present the proposed algorithm, MSPSA, and its theoretical learning guarantees for each of the objectives considered separately. Then, we explain how the dynamic pricing problem of an electricity retailer maps to the online learning problem studied here. We conclude the chapter with simulation study to illustrate the regret growth and convergence of MSPSA algorithm.

2.1 Problem Formulation

The model considered here is an affine model, modulated by an exogenous finite state time-homogeneous Markov chain (\mathcal{S}, P) where $\mathcal{S} = \{1, \dots, K\}$ is the state space and $P = [p_{i,j}]$ the transition probability matrix. We assume that the state space \mathcal{S} and the transition matrix P are unknown.

Each state $k \in \mathcal{S}$ of the Markov chain is associated with an affine model whose parameters are denoted by $\theta_k = (A_k, b_k)$ where $A_k \in \mathbb{R}^{m \times n}$ has full column rank and $b_k \in \mathbb{R}^m$. All system parameters $\theta = \{\theta_k\}_{k=1}^K$ are assumed deterministic and unknown. At time t , the input-output relation of the system is given by

$$y_t = A_{s_t} x_t + b_{s_t} + w_t, \quad (2.1)$$

where $x_t \in \mathbb{R}^n$ is the control input, $y_t \in \mathbb{R}^m$ the observable output, $s_t \in \mathcal{S}$ the state of the system, and $w_t \in \mathbb{R}^m$ is a random vector that captures the system

noise. It is assumed that the random noise w_t is drawn from a possibly state dependent distribution $f_{s_t}(\cdot)$ with zero mean (without loss of generality) and unknown finite variance $\Sigma_w^{(s_t)}$. Furthermore, for any $t \neq t'$, w_t and $w_{t'}$ are conditionally independent given the states s_t and $s_{t'}$.

Before choosing the control input of period t , the only information the decision maker has is a vector I_{t-1} containing its decision and observation history up to time $t-1$, which consists of input vector $X^{t-1} = (x_1, \dots, x_{t-1})$, output vector $Y^{t-1} = (y_1, \dots, y_{t-1})$, state vector $S^{t-1} = (s_0, \dots, s_{t-1})$, and a convex compact set $\Pi \in \mathbb{R}^n$ containing the optimal input (solution) under known model (which is explained later on in this section).

Even though the assumption on the Markov process being observable and exogenous is restrictive, there are ample applications that can be well modeled/approximated by observable and exogenous Markov dynamics. Later, we will present the dynamic pricing problem of an electricity retailer in detail. Other relevant applications include data transmission under changing channel conditions that are exogenous to the message transmitted [83] and control of an unmanned aerial vehicle where exogenous conditions such as wind affect the underlying model [5].

The objective of *online learning and optimization* is to find a control input sequence $\{x_t\}_{t=1}^T$ that minimizes the expected cumulative cost incurred at each stage. The stage cost at time t is defined as the expected cost incurred at time t where the expectation is conditioned on the current observation history I_{t-1} and the control input x_t . Due to Markov chain, however, the stage cost $\mathcal{J}(s_{t-1}, x_t)$ at time t becomes a function of the most recently observed state s_{t-1} and the

control input x_t only.* Then, the expected cumulative cost can be expressed as

$$\mathbb{E} \left(\sum_{t=1}^T \mathcal{J}(s_{t-1}, x_t) \right),$$

where T is the learning and optimization horizon. Note that the above quantity is a function of the deterministic parameters θ and the distributions P and $\{f_i\}_{i=1}^K$.

For a decision maker who wants to minimize its expected cumulative cost, the difficulty in finding the optimal control input sequence is that the system parameter θ and the transition matrix P are unknown. If the system parameters (θ, P) were known, then the decision maker would have used this information along with its observation history to determine the optimal decision rule that minimizes the expected cumulative cost. In that case, the problem could be formulated as a dynamic program and solved via backward induction. We refer to this optimal solution under known model as the optimal input and denote it by $\{x_t^*\}_{t=1}^T$ (which is made precise in the following sections). As mentioned before, the optimal input x_t^* is contained in the set Π , which is known to the decision maker.

A policy μ of a decision maker is defined as a sequence of decision rules, *i.e.*, $\mu = (\mu_0, \mu_1, \dots, \mu_{T-1})$, such that, at time $t-1$, μ_{t-1} maps the information history vector I_{t-1} to the system input x_t at time t . We denote the input determined by policy μ as x_t^μ .

The online learning of Markov jump affine model is depicted in Figure 2.1 by a feedback loop. The unknown variables are indicated in red in the figure. Here, one can observe how the decision of a policy effects its observation and

*This can be observed in (2.3), the stage cost for quadratic regulation problem, and in (2.4), the stage costs for revenue maximization problem.

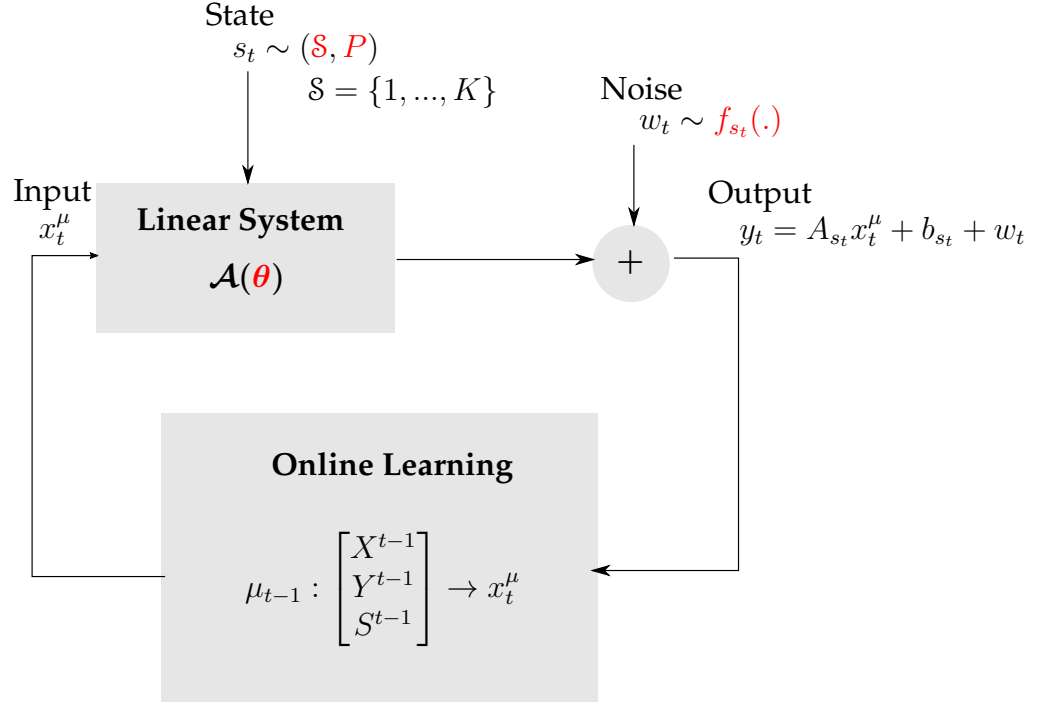


Figure 2.1: Online learning of Markov jump affine model

vice versa. The figure also illustrates the idea behind the exploration and exploitation tradeoff. In order to determine the optimal input accurately, it is necessary to make better predictions regarding the unknown variables. However, in order to make better predictions, it is crucial to explore the space of possible inputs to get observations that provide necessary information. So, what you can exploit is determined by how much you explore. On the other hand, exploring the space of possible inputs restricts how much you can exploit. Hence, the key is to find the policy μ that establishes the optimal balance between exploration and exploitation.

To measure the performance of an online learning policy, we use the regret measure as a proxy. In particular, the (cumulative) regret $\mathcal{R}_T^\mu(\theta, P)$ of a learning policy μ is measured by the difference between the expected cumulative cost of the decision maker, who follows the policy μ , and that of a decision maker who

knows the system parameters (θ, P) and sets the system input optimally, *i.e.*,

$$\mathcal{R}_T^\mu(\theta, P) = \mathbb{E} \left(\sum_{t=1}^T \mathcal{J}(s_{t-1}, x_t^\mu) - \sum_{t=1}^T \mathcal{J}(s_{t-1}, x_t^*) \right). \quad (2.2)$$

Since the regret defined above is a function of system parameters, we characterize the performance of μ by the worst case regret

$$\bar{\mathcal{R}}_T^\mu \triangleq \sup_{\theta, P, \{f_i\}_{i=1}^K} \mathcal{R}_T^\mu(\theta, P).$$

Here we are interested in the worst case (but non-adversarial) parameter θ in a compact set $\Theta \subset \mathbb{R}^{K \times m \times (n+1)}$, independent of the learning horizon T . For the worst case analysis, it is assumed that the state space size K , and system dimensions m and n are fixed. Since Θ is compact, for any $\theta \in \Theta$ and for all $k \in \mathcal{S}$, the largest singular value of A_k is bounded by a positive constant $\bar{\sigma}$, and the parameter b_k is bounded by a positive constant \bar{b} , *i.e.*, $\|b_k\|_2 \leq \bar{b}$. It is also assumed that the variance of w_t is bounded, *i.e.*, $\mathbb{E}(w_{t,i}^2) \leq \sigma_w^2$ for some positive constant σ_w where $w_{t,i}$ denotes the i th entry of w_t . The optimal input for the worst-case parameters (θ, P) certainly has to be contained in the set Π , and A_k has to be full column rank for every $k \in \mathcal{S}$ for the worst-case parameter θ . Note that $\bar{\mathcal{R}}_T^\mu$ grows monotonically with T . We are interested in the learning rule that has the slowest regret growth.

In the following sections, we focus on two different stage costs, hence two different objective functions. The first is a quadratic cost that arises naturally from the regulation problem. In particular, the stage cost at time t is given by

$$\mathcal{J}^Q(s_{t-1}, x_t) \triangleq \mathbb{E}(\|y^* - y_t\|_2^2 | s_{t-1}, x_t), \quad (2.3)$$

where $y^* \in \mathbb{R}^m$ is a constant target value for output. For the quadratic regulation

problem, we assume that the forth order moment of w_t is bounded, *i.e.*, $\mathbb{E}(w_{t,i}^4) \leq \sigma_w^4$, in addition to the previous boundedness assumptions.

The second stage cost we consider is the minus revenue that arises from revenue maximization problem (or profit maximization problem since profit can be expressed as revenue minus total cost). Specifically, the stage cost of period t is given by

$$\mathcal{J}^R(s_{t-1}, x_t) \triangleq \mathbb{E}(-x_t^\top y_t | s_{t-1}, x_t). \quad (2.4)$$

Here, the revenue is calculated as the inner product of the input and the output vector where the entries of the input and the output vector corresponds to the price and the demand of each product, respectively. Therefore, the input and the output dimensions match, *i.e.*, $m = n$, for the revenue maximization problem. For this objective, it is assumed that the matrix A_j is negative definite for all $j \in \mathcal{S}$ which is a reasonable assumption in dynamic pricing problems, *e.g.*, see [39,42].

2.2 Online Learning for Quadratic Regulation

In this section, we study the online learning problem with the quadratic cost. We first derive an expression for regret using the optimal solution under known model referred to as the optimal input. Then, we introduce an online learning approach and establish its order optimality via the analysis of its regret growth rate and the analysis of the minimum regret growth rate achievable by any policy. We also show that the input of the online learning policy converges to the optimal input both almost surely and in mean square.

2.2.1 Optimal Solution Under Known Model and Regret

In order to calculate the regret for any policy, we begin by deriving the optimal solution of a decision maker who knows the system (θ, P) in addition to I_{t-1} and aims to minimize the expected cumulative cost, *i.e.*,

$$\min_{\{x_t\}_{t=1}^T} \mathbb{E} \left(\sum_{t=1}^T \mathcal{J}^Q(s_{t-1}, x_t) \right). \quad (2.5)$$

The problem under known model becomes a dynamic program due to the known (θ, P) . Since the Markov process is exogenous, *i.e.*, independent of the decision policy, the optimization problem decouples to choosing the system input x_t separately for each decision stage with stage cost given in (2.3) which is equivalent to

$$\mathcal{J}^Q(s_{t-1}, x_t) = \sum_j p_{s_{t-1},j} (\|y^* - A_j x_t - b_j\|_2^2 + \text{tr}(\Sigma_w^{(j)})) \quad (2.6)$$

by (2.1). The optimal input x_t^* minimizing the stage cost is then given by

$$x_{s_{t-1}}^* = \left(\sum_j p_{s_{t-1},j} A_j^\top A_j \right)^{-1} \left(\sum_j p_{s_{t-1},j} A_j^\top (y^* - b_j) \right). \quad (2.7)$$

Thus, the optimal input $x_t^* \in \Pi$ at any time t depends only on the system parameter θ , the transition matrix P , and the previous state s_{t-1} . In the sequel, we use $x_{s_{t-1}}^*$ to represent x_t^* , dropping the explicit parameter dependency on (θ, P) in the notation.

Hence, the stage regret at t , which is the expected difference of the stage cost obtained by policy μ and the stage cost of the optimal input $x_{s_{t-1}}^*$, can be written as

$$\begin{aligned} r_t^\mu(\theta, P) &= \mathbb{E} \left(\mathcal{J}^Q(s_{t-1}, x_t^\mu) - \mathcal{J}^Q(s_{t-1}, x_{s_{t-1}}^*) \right) \\ &= \mathbb{E} \left(\|A_{s_t}(x_t^\mu - x_{s_{t-1}}^*)\|_2^2 \right), \end{aligned}$$

which is obtained using the first order optimality condition (FOC) for $x_{s_{t-1}}^*$. The T-period regret given in (2.2) can then be expressed as

$$\mathcal{R}_T^\mu(\theta, P) = \mathbb{E} \left(\sum_{t=1}^T \|A_{s_t}(x_t^\mu - x_{s_{t-1}}^*)\|_2^2 \right). \quad (2.8)$$

2.2.2 MSPSA: An Online Learning Policy

Here, we present an online learning policy to the quadratic regulation problem that achieves the slowest regret growth rate possible. Referred to as MSPSA, the policy is an extension of the simultaneous perturbation stochastic approximation (SPSA) algorithm proposed by Spall [74] to Markov jump models considered here.

Spall's SPSA is a stochastic approximation algorithm that updates the estimate of the optimal input by a stochastic approximation of the objective gradient. The key step is to generate two consecutive observations corresponding to two inputs, that are set to be the current optimal-input estimate perturbed by some random vector in opposite directions, and use them to construct the gradient estimate. In applying this idea to the optimization problem of a Markov jump system, a complication arises due to the uncertainty associated with the system state at the time when the system input is determined; consecutive observations that are used to determine the gradient estimate may correspond to different system states.

The key idea of MSPSA is to keep track of each state $i \in \mathcal{S}$ and the estimate of the optimal input associated with each state $i \in \mathcal{S}$. When state i is realized, the estimate of the optimal input associated with state i is perturbed by some random vector and this randomly perturbed estimate is used as input for the


```

1: for  $t = 1$  to  $T$  do
2:   if  $s_{t-1} = i$  is observed then
3:     if state  $i$  is observed for the first time then
4:       Let  $\hat{x}_{i,1} \in \Pi$  be an arbitrary vector
5:        $t_i \leftarrow 0$ 
6:        $e_i \leftarrow 0$ 
7:     end if
8:     if  $e_i = 0$  then
9:        $t_i \leftarrow t_i + 1$ 
10:       $x_t \leftarrow \hat{x}_{i,t_i} + c_{t_i} \Delta_{t_i}$ 
      where  $c_{t_i} = \gamma'_i / (N'_i + t_i)^{0.25}$  with some positive constant  $\gamma'_i$  and a non-
      negative integer  $N'_i$ , and  $\Delta_{t_i} = [\Delta_{t_i,1}, \dots, \Delta_{t_i,n}]^\top$  with  $\Delta_{t_i,j}$ 's drawn from an
      independent and identical distribution that is symmetrical around zero,
      and satisfies  $|\Delta_{t_i,j}| \leq \xi_1$  and  $\mathbb{E}(1/\Delta_{t_i,j}^2) \leq \xi_2$  for some positive constants  $\xi_1$ 
      and  $\xi_2$ .
11:       $d_{i,t_i}^+ \leftarrow \|y_t - y^*\|_2^2$ 
12:       $e_i \leftarrow 1$ 
13:    else
14:       $x_t \leftarrow \hat{x}_{i,t_i} - c_{t_i} \Delta_{t_i}$ 
15:       $d_{i,t_i}^- \leftarrow \|y_t - y^*\|_2^2$ 
16:       $e_i \leftarrow 0$ 
17:    Update:

$$\hat{x}_{i,t_i+1} \leftarrow \left( \hat{x}_{i,t_i} - a_{t_i} \left( \frac{d_{i,t_i}^+ - d_{i,t_i}^-}{c_{t_i}} \right) \bar{\Delta}_{t_i} \right)_\Pi \quad (2.9)$$

      where  $(\cdot)_\Pi$  denotes the euclidean projection operator onto  $\Pi$ ,  $\bar{\Delta}_{t_i} =$ 
 $[1/\Delta_{t_i,1}, \dots, 1/\Delta_{t_i,n}]^\top$ , and  $a_{t_i} = \gamma_i / (N_i + t_i)$  with some positive constant  $\gamma_i$ 
      and a non-negative integer  $N_i$ 
18:    end if
19:  end if
20: end for

```

Figure 2.2: MSPSA algorithm pseudocode for quadratic regulation problem

next stage. The estimate of the optimal input associated with state i is updated only when we obtain two observations of the system output corresponding to two inputs that are generated by perturbing the current estimate in opposite directions by the same amount right after observing state i .

Details of this implementation is given in Figure 2.2. Whenever a new state $i \in \mathcal{S}$ is observed that has not been observed before, MSPSA policy assigns an arbitrary predetermined vector $\hat{x}_{i,1} \in \Pi$ as the initial estimate of the optimal input x_i^* (line 3-7 of Figure 2.2). At the beginning of each stage t , MSPSA checks the previous state s_{t-1} (line 2), and whether any observation is taken using the most recent optimal-input estimate $\hat{x}_{s_{t-1}, t_{s_{t-1}}}$ (line 8) where $t_{s_{t-1}}$ is the number of times the optimal-input estimate $\hat{x}_{s_{t-1}, t_{s_{t-1}}}$ is updated up to time t (Since two observations, that are taken right after observing state s_{t-1} , are used for each update of $\hat{x}_{s_{t-1}, t_{s_{t-1}}}$, $t_{s_{t-1}}$ is approximately half of the number of times state s_{t-1} is observed up to t). If an observation has not taken using the most recent estimate yet, the input for that stage is set to be a randomly perturbed $\hat{x}_{s_{t-1}, t_{s_{t-1}}}$ (line 10). Otherwise MSPSA sets the input by perturbing the estimate $\hat{x}_{s_{t-1}, t_{s_{t-1}}}$ in the opposite direction by the same amount as the previous one (line 14). Then, it updates the optimal-input estimate by a stochastic approximation (line 17) obtained using the stage costs calculated from both observations (line 11 and 15) and projects it onto Π . The constant γ'_i of the perturbation gain sequence c_{t_i} should be chosen larger in the high noise setting for an accurate gradient estimate. The choice of the sequence a_{t_i} used for the update step determines the step size. The non-negative integers N_i of a_{t_i} and N'_i of c_{t_i} can be set to zero as default, but if the update of the optimal-input estimate fluctuates between the borders of Π at the beginning of the MSPSA policy, setting N_i greater than zero can prevent this fluctuation.

2.2.3 Regret Analysis for MSPSA

With MSPSA, we present the idea of state tracking that is applied to SPSA algorithm to deal with the Markov jump dynamics. Even though applying the idea of state tracking is an intuitive extension, it is not obvious if this type of extension of existing methods that are used for time-invariant case would actually converge to the optimal solution. For example, when the same idea of state tracking is applied to Robbins-Monro algorithm, (which has optimal regret growth for the time invariant setting of the quadratic regulation problem with negative definite system matrix) the solution converges to a sub-optimal point resulting in a linear regret growth. Yet, we show that it is possible to achieve regret growth rate of $O(\sqrt{T})$ with MSPSA that introduces the idea of state tracking applied to SPSA algorithm. We also show that the input generated by MSPSA converges to the optimal solution both almost surely and in mean square.

We now analyze the regret performance of MSPSA. Let $\lambda_{\min}(\cdot)$ denote the minimum eigenvalue operator, and $e_{i,t_i} = \mathbb{E}(\|\hat{x}_{i,t_i} - x_i^*\|_2^2 | i, t_i)$ be the mean squared error (MSE) between the optimal input x_i^* and its estimate \hat{x}_{i,t_i} given state i and t_i , where t_i , as defined in previous section, is the number of times the estimate \hat{x}_{i,t_i} has been updated up to time t by MSPSA. The following lemma provides a bound for the decreasing rate of e_{i,t_i} , and hence the convergence rate of the estimate to its true value in terms of the number of times the estimate is updated and thus in terms of the number of times the state i has occurred up to t (which is equal either to $2t_i$ or to $2t_i - 1$). It shows that the MSE converges to zero with a rate equal or faster than the inverse of the square root of the number of times state i has occurred.

Lemma 1. *For any $i \in \mathcal{S}$, if $\gamma_i \geq 1/(8\lambda_{\min}(\sum_j p_{i,j} A_j^\top A_j))$ then there exists a constant*

$C_i > 0$ satisfying $e_{i,t_i} \leq C_i/\sqrt{t_i}$ for any $(\theta, P, \{f_i\}_{i=1}^K)$.

Proof. See Appendix. □

To satisfy the condition of Lemma 1, the decision maker, who follows MSPSA, needs to have some information about a lower bound on the minimum eigenvalue of $\sum_j p_{i,j} A_j^\top A_j$, e.g., knowing a non-trivial lower bound $\underline{\sigma}$ for the singular values of the system matrices. This assumption may not be restrictive in practice since the decision maker can set γ_i sufficiently large.

Let the worst-case cumulative input-MSE $\bar{\mathcal{E}}_T^\mu$ be the worst-case cumulative MSE between the input x_t^μ of policy μ and the optimal input $x_{s_{t-1}}^*$, i.e.,

$$\bar{\mathcal{E}}_T^\mu \triangleq \sup_{\theta, P, \{f_i\}_{i=1}^K} \mathbb{E} \left(\sum_{t=1}^T \|x_t^\mu - x_{s_{t-1}}^*\|_2^2 \right).$$

Using the result of Lemma 1, we provide a bound for the growth rate of the worst-case cumulative input-MSE and the worst-case regret of MSPSA. Theorem 1 shows that the MSPSA policy achieves the regret growth rate of $O(\sqrt{T})$.

Theorem 1. *If $\gamma_i \geq 1/(8\lambda_{\min}(\sum_j p_{i,j} A_j^\top A_j))$ for every $i \in \mathcal{S}$, then there exist some positive constants C and C' such that*

$$\bar{\mathcal{E}}_T^{\text{MSPSA}} \leq C\sqrt{T},$$

and

$$\bar{\mathcal{R}}_T^{\text{MSPSA}} \leq C'\sqrt{T}.$$

Proof. The input of MSPSA x_t^{MSPSA} is equal to either $\hat{x}_{i,t_i} + c_{t_i}\Delta_{t_i}$ or $\hat{x}_{i,t_i} - c_{t_i}\Delta_{t_i}$ given $s_{t-1} = i$ and t_i . By Lemma 1, observe that

$$\begin{aligned} r_{i,t_i} &= \mathbb{E} \left(\|\hat{x}_{i,t_i} \pm c_{t_i}\Delta_{t_i} - x_i^*\|_2^2 | i, t_i \right) \\ &= e_{i,t_i} + c_{t_i}^2 \mathbb{E}(\Delta_{t_i}^\top \Delta_{t_i}) \leq C'_i/\sqrt{t_i} \end{aligned} \tag{2.10}$$

where $C'_i = C_i + (\gamma'_i)^2 n \xi_1^2$. Let T_i be the number of times the estimate of the optimal input associated with state i has been updated until period T . Because MSPSA uses two observations per update, we can express the worst-case cumulative input-MSE for MSPSA as

$$\bar{\mathcal{E}}_T^{\text{MSPSA}} = \sup_{\theta, P, \{f_i\}_{i=1}^K} \mathbb{E} \left(\sum_{i=1}^K \sum_{t_i=1}^{T_i} 2r_{i,t_i} \right).$$

By (2.10), we bound $\sum_{t_i=1}^{T_i} 2r_{i,t_i} \leq C_0 \sqrt{T_i}$ where $C_0 = 4 \max_{i \in \mathcal{S}} C'_i$. Since T_i is smaller than the number of times state i is observed, which is a fraction of T for any P , $\bar{\mathcal{E}}_T^{\text{MSPSA}} \leq C \sqrt{T}$ where $C = \sqrt{K} C_0$. Consequently, by (2.8) and using the upper bound $\bar{\sigma}$ for the singular values of A_{s_t} , $\bar{\mathcal{R}}_T^{\text{MSPSA}} \leq \bar{\sigma}^2 \bar{\mathcal{E}}_T^{\text{MSPSA}} \leq C' \sqrt{T}$ where $C' = \bar{\sigma}^2 C$. \square

According to Theorem 1, the average regret converges to zero with a rate equal or faster than $1/\sqrt{T}$. Hence, it proves that the average performance of MSPSA policy approaches to that of the optimal solution under known model as $T \rightarrow \infty$. However, this does not imply the convergence of the MSPSA policy to the optimal input. The following theorem provides both almost sure and mean square convergence of MSPSA policy to the optimal input.

Theorem 2. *For the quadratic regulation problem,*

$$\Pr \left(\lim_{T \rightarrow \infty} \|x_T^{\text{MSPSA}} - x_{s_{T-1}}^*\|_2^2 = 0 \right) = 1, \quad (2.11)$$

and

$$\lim_{T \rightarrow \infty} \mathbb{E} \left(\|x_T^{\text{MSPSA}} - x_{s_{T-1}}^*\|_2^2 \right) = 0. \quad (2.12)$$

Proof. See Appendix. \square

Theorem 2 shows that, the input generated by MSPSA converges to its optimal value as $T \rightarrow \infty$ for any choice of $\gamma_i > 0$. Hence, the condition given

in Theorem 1 is not necessary for the convergence of MSPSA policy. The intuition is as follows; for any observed recurrent state $i \in \mathcal{S}$, the input of MSPSA converges to its optimal value at time periods when the previous state is i as $T \rightarrow \infty$, and any observed transient state $i \in \mathcal{S}$ will occur only a finite number of times. Therefore, the input of MSPSA converges to its optimal value as $T \rightarrow \infty$.

2.2.4 A Lower Bound on the Growth Rate of Regret

We now show that MSPSA in fact provides the slowest possible regret growth. To this end, we provide a lower bound of regret growth for all decision policies.

For any policy μ , the estimate of the optimal input and the actual input of the policy may not be the same; for example, the input of MSPSA is a randomly perturbed estimate of the optimal input and not the estimate itself. Hence, let's denote an optimal-input estimate obtained using the past observations corresponding to inputs of policy μ at time t by \hat{x}_t^μ and the input at time t by x_t^μ . We define the worst-case cumulative estimation-MSE as

$$\hat{\mathcal{E}}_T^\mu \triangleq \sup_{\theta, P, \{f_i\}_{i=1}^K} \mathbb{E} \left(\sum_{t=1}^T \|\hat{x}_t^\mu - x_{s_{t-1}}^*\|_2^2 \right).$$

In particular, the following theorem states that the product of the growth rate of the worst-case cumulative input-MSE $\bar{\mathcal{E}}_T^\mu$ and the worst-case cumulative estimation-MSE $\hat{\mathcal{E}}_T^\mu$ of any sequence $\{\hat{x}_t^\mu\}_{t=1}^T$ cannot be lower than T for any policy μ .

Theorem 3. *For any value of $K > 1$, there exists a constant $C > 0$ such that, for any*

policy μ ,

$$\hat{\mathcal{E}}_T^\mu \bar{\mathcal{E}}_T^\mu \geq CT. \quad (2.13)$$

Proof. See Appendix. □

Theorem 3 shows the trade-off between exploration (minimizing the estimation error) and exploitation (minimizing the input error): the product of the cumulative estimation-MSE and the cumulative input-MSE grows linearly with T for any policy. This result implies that if the goal is to minimize the cumulative estimation-MSE rather than the regret, then it is possible to find a policy for which $\hat{\mathcal{E}}_T^\mu$ grows slower than \sqrt{T} in which case the cumulative input-MSE $\bar{\mathcal{E}}_T^\mu$ has to grow faster than \sqrt{T} . In fact, if the perturbation gain sequence c_{t_i} is set to be constant rather than a decreasing sequence of t_i , by following the proof of Theorem 1, it is easy to show that MSPSA's cumulative estimation-MSE grows no faster than $\log T$ whereas its regret would grow linearly with T .

However, the slowest growth rate of $\bar{\mathcal{E}}_T^\mu$ cannot be slower than that of $\hat{\mathcal{E}}_T^\mu$ for the optimal choice of the estimate sequence $\{\hat{x}_t^\mu\}_{t=1}^T$ (in other words, one can always take the estimate equal to the input, *i.e.*, $\hat{x}_t^\mu = x_t^\mu$, in which case $\hat{\mathcal{E}}_T^\mu = \bar{\mathcal{E}}_T^\mu$). Therefore, the growth rate of the worst-case cumulative input-MSE $\bar{\mathcal{E}}_T^\mu$, and, consequently, the growth rate of the worst-case regret cannot be lower than \sqrt{T} for any policy μ . Hence, the regret growth rate of MSPSA is the optimal one and achieves the lower bound $\Omega(\sqrt{T})$, which is stated in Theorem 4.

Theorem 4. *For any value of $K > 1$, there exist some constants $C', C'' > 0$ such that, for any policy μ ,*

$$\bar{\mathcal{E}}_T^\mu \geq C'\sqrt{T}, \quad (2.14)$$

and

$$\bar{\mathcal{R}}_T^\mu \geq C''\sqrt{T}. \quad (2.15)$$

Proof. We choose the estimate \hat{x}_t^μ equal to the input x_t^μ . Then, by Theorem 3, we have $(\bar{\mathcal{E}}_T^\mu)^2 \geq CT$. As a result, $C' = \sqrt{C}$. Let $\bar{\theta} \in \Theta$ be the parameter satisfying $\mathbb{E}(\sum_{t=1}^T \|x_t^\mu - x_{s_{t-1}}^*\|_2^2 | \bar{\theta}) = \bar{\mathcal{E}}_T^\mu$. In the proof of Theorem 3, we fixed $p_{i,j} = 1/K$ for any $i, j \in \mathcal{S}$. Hence, by (2.8), $\mathcal{R}_T^\mu(\bar{\theta}, P) \geq \underline{\sigma} \bar{\mathcal{E}}_T^\mu$ where $\underline{\sigma} = \min_{\theta \in \Theta} \lambda_{\min}(\sum_j A_j^\top A_j / K) > 0$ by the extreme value theorem. Then, the worst case regret $\bar{\mathcal{R}}_T^\mu \geq \mathcal{R}_T^\mu(\bar{\theta}, P) \geq C''\sqrt{T}$ where $C'' = \underline{\sigma}C'$. \square

To prove Theorem 3, we consider a hypothetical case in which the decision maker receives additional observations at each period t . It is assumed that the additional observations provided to the decision maker are the observation values corresponding to input x_t^μ from the states that didn't occur at t . Since such observations can't increase the growth rate of regret of the optimal policy, we establish a lower bound for this case by showing that it becomes equivalent to a single state case with $m > n$ and using the multivariate van Trees inequality [35] in a similar way as in [42]. If $K = 1$ and $m > n$, the proofs of Theorem 3 and Theorem 4 lead to the result regarding the single state case given in Corollary 1.

Corollary 1. *For $K = 1$ and for any value of m and n satisfying $m > n$, there exist some constants $C, C', C'' > 0$ such that, for any policy μ , inequalities (2.13), (2.14), and (2.15) hold.*

As mentioned in related work, it has been shown that for $K = 1$ and $m = n$ case, the regret growth rate is $\Theta(\log T)$ [51]. We show that the characteristics of regret growth changes from $\Theta(\log T)$ to $\Theta(\sqrt{T})$ for Markov jump system. Additionally, Corollary 1 states that, even in the absence of Markov jump, when

system matrix A is not invertible, the best regret growth rate also jumps from $\Theta(\log T)$ to $\Theta(\sqrt{T})$. The most significant difference from the single state case with invertible system matrix is that the cost function for Markov jump system or for single state system with $m > n$ given in (2.6) is not the root of the cost function as in the case of $K = 1$ with $m = n$, and decision maker can't understand how close it is to the minimum just by looking at its observations.

2.3 Online Learning for Revenue Maximization

The single state setting of the revenue maximization problem has been previously studied by Keskin and Zeevi [42]. Here, we consider the more general setting where the affine demand parameters can change depending on the state of nature, more precisely the setting where demand parameters follows a Markov jump process.

As for the quadratic regulation objective, to obtain a regret expression for revenue maximization objective, we first determine the optimal solution of a decision maker who knows the system (θ, P) . Then, we present MSPSA policy for revenue maximization problem and establish its optimality in regret performance and its convergence to the optimal solution.

2.3.1 Optimal Solution Under Known Model and Regret

By following the same argument as before, under known model, the optimal solution of a decision maker aimed at minimizing the expected cumulative cost given in (2.5), which is equal to minus expected T-period revenue, is to choose

the system input minimizing the respective stage cost given in (2.4). By using (2.1), this stage cost can be written as

$$\mathcal{J}^R(s_{t-1}, x_t) = - \sum_j p_{s_{t-1},j} x_t^\top (A_j x_t + b_j). \quad (2.16)$$

The optimal input x_t^* , which depends only on (θ, P) and the previous state s_{t-1} , is then given by

$$x_{s_{t-1}}^* = - \left(\sum_j p_{s_{t-1},j} (A_j + A_j^\top) \right)^{-1} \left(\sum_j p_{s_{t-1},j} b_j \right),$$

by dropping the explicit dependency of x_t^* on (θ, P) in the notation.

Using the FOC for the optimal input $x_{s_{t-1}}^*$, we obtain the stage regret of a policy μ , i.e.,

$$\begin{aligned} r_t^\mu(\theta, P) &= \mathbb{E} \left(\mathcal{J}^R(s_{t-1}, x_t^\mu) - \mathcal{J}^R(s_{t-1}, x_{s_{t-1}}^*) \right) \\ &= -\mathbb{E} \left((x_t^\mu - x_{s_{t-1}}^*)^\top A_{s_t} (x_t^\mu - x_{s_{t-1}}^*) \right). \end{aligned}$$

Since A_{s_t} is negative definite, the stage regret is always non-negative. Consequently, the T-period regret is given by

$$\mathcal{R}_T^\mu(\theta, P) = -\mathbb{E} \left(\sum_{t=1}^T (x_t^\mu - x_{s_{t-1}}^*)^\top A_{s_t} (x_t^\mu - x_{s_{t-1}}^*) \right). \quad (2.17)$$

2.3.2 MSPSA Policy for Revenue Maximization

Here, we present the MSPSA policy for revenue maximization objective. The only difference between the two problems considered is their respective stage costs (objectives). Therefore, the only change in MSPSA policy is how the stage costs are calculated to approximate the objective gradient which corresponds to line 11 and 15 of Figure 2.2. In the corresponding steps of MSPSA policy for revenue maximization, that is given in Figure 2.3 in details, the stage costs are calculated as minus the observed revenue at that stage.

```

1: for  $t = 1$  to  $T$  do
2:   if  $s_{t-1} = i$  is observed then
3:     if state  $i$  is observed for the first time then
4:       Let  $\hat{x}_{i,1} \in \Pi$  be an arbitrary vector
5:        $t_i \leftarrow 0$ 
6:        $e_i \leftarrow 0$ 
7:     end if
8:     if  $e_i = 0$  then
9:        $t_i \leftarrow t_i + 1$ 
10:       $x_t \leftarrow \hat{x}_{i,t_i} + c_{t_i} \Delta_{t_i}$ 
      where  $c_{t_i} = \gamma'_i / (N'_i + t_i)^{0.25}$  with some positive constant  $\gamma'_i$  and a non-
      negative integer  $N'_i$ , and  $\Delta_{t_i} = [\Delta_{t_i,1}, \dots, \Delta_{t_i,n}]^\top$  with  $\Delta_{t_i,j}$ 's drawn from an
      independent and identical distribution that is symmetrical around zero,
      and satisfies  $|\Delta_{t_i,j}| \leq \xi_1$  and  $\mathbb{E}(1/\Delta_{t_i,j}^2) \leq \xi_2$  for some positive constants  $\xi_1$ 
      and  $\xi_2$ .
11:       $d_{i,t_i}^+ \leftarrow -x_t^\top y_t$ 
12:       $e_i \leftarrow 1$ 
13:    else
14:       $x_t \leftarrow \hat{x}_{i,t_i} - c_{t_i} \Delta_{t_i}$ 
15:       $d_{i,t_i}^- \leftarrow -x_t^\top y_t$ 
16:       $e_i \leftarrow 0$ 
17:    Update:

$$\hat{x}_{i,t_i+1} \leftarrow \left( \hat{x}_{i,t_i} - a_{t_i} \left( \frac{d_{i,t_i}^+ - d_{i,t_i}^-}{c_{t_i}} \right) \bar{\Delta}_{t_i} \right)_\Pi \quad (2.18)$$

      where  $(\cdot)_\Pi$  denotes the euclidean projection operator onto  $\Pi$ ,  $\bar{\Delta}_{t_i} =$ 
 $[1/\Delta_{t_i,1}, \dots, 1/\Delta_{t_i,n}]^\top$ , and  $a_{t_i} = \gamma_i / (N_i + t_i)$  with some positive constant  $\gamma_i$ 
      and a non-negative integer  $N_i$ .
18:    end if
19:  end if
20: end for

```

Figure 2.3: MSPSA algorithm pseudocode for revenue maximization problem

2.3.3 MSPSA Performance and Regret Lower Bound

To obtain the regret growth rate for MSPSA policy for revenue maximization, we first derive an upper bound on how fast the estimate \hat{x}_{i,t_i} of the optimal input x_i^* converges to its true value as we did for the regulation problem. Lemma 2 shows that the conditional MSE e_{i,t_i} between the optimal input and its estimate converges to zero with a rate no smaller than the inverse of the square root of the number of times the estimate \hat{x}_{i,t_i} is updated by MSPSA.

Lemma 2. *For any $i \in \mathcal{S}$, if $\gamma_i \geq 1/(8\lambda_{\min}(-\sum_j p_{i,j}(A_j + A_j^\top)/2))$ then there exists a constant $C_i > 0$ satisfying $e_{i,t_i} \leq C_i/\sqrt{t_i}$ for any $(\theta, P, \{f_i\}_{i=1}^K)$.*

Proof. See Appendix. □

The condition of Lemma 2 is slightly different than that of Lemma 1. The bound on the step size constant γ_i depends on the minimum eigenvalue of $-\sum_j p_{i,j}(A_j + A_j^\top)/2$. This difference is due to the choice of a different stage cost. However, the information of a non-trivial lower bound $\underline{\sigma}$ on the singular values of the system matrices is still sufficient to satisfy this condition.

Using the result of Lemma 2, we prove that MSPSA achieves the regret growth rate of $O(\sqrt{T})$ for revenue maximization objective as given in Theorem 5, and the input of MSPSA converges to the optimal input both almost surely and in mean square as given in Theorem 6.

Theorem 5. *If $\gamma_i \geq 1/(8\lambda_{\min}(-\sum_j p_{i,j}(A_j + A_j^\top)/2))$ for every $i \in \mathcal{S}$, then there exist some positive constants C and C' such that*

$$\bar{\mathcal{E}}_T^{\text{MSPSA}} \leq C\sqrt{T},$$

and

$$\bar{\mathcal{R}}_T^{\text{MSPSA}} \leq C' \sqrt{T}.$$

Proof. Same as the proof of Theorem 1 up to the step that $\bar{\mathcal{E}}_T^{\text{MSPSA}} \leq C\sqrt{T}$ is obtained. Then, by the regret given in (2.17) for revenue maximization objective and the fact that $-(A_j + A_j^\top)/2$ is positive definite with eigenvalues upper bounded by $\bar{\sigma}$, $\bar{\mathcal{R}}_T^{\text{MSPSA}} \leq \bar{\sigma} \bar{\mathcal{E}}_T^{\text{MSPSA}} \leq C' \sqrt{T}$ where $C' = \bar{\sigma}C$. \square

Theorem 6. *For revenue maximization problem,*

$$\Pr \left(\lim_{T \rightarrow \infty} \|x_T^{\text{MSPSA}} - x_{s_{T-1}}^*\|_2^2 = 0 \right) = 1,$$

and

$$\lim_{T \rightarrow \infty} \mathbb{E} \left(\|x_T^{\text{MSPSA}} - x_{s_{T-1}}^*\|_2^2 \right) = 0.$$

Proof. In the proof of Lemma 2, we showed that (A.4) holds for any state $i \in \mathcal{S}$. Therefore, the proof follows the proof of Theorem 2. \square

Previously, for single state setting of this problem, Keskin and Zeevi [42] have shown that for any policy μ the worst case regret growth for this problem cannot be smaller than $\Omega(\sqrt{T})$ and they have shown that $O(\sqrt{T} \log T)$ is achievable by a semi-myopic policy that they referred to as multivariate constrained iterated least squares (MCILS) policy. Here, we showed that it is possible to achieve the lower bound $\Omega(\sqrt{T})$ given in [42] by MSPSA policy for more general problem with Markov jumped demand.

Next, we generalize Keskin and Zeevi's lower bound result to Markov jump case by showing that for any policy μ and for any state space size $K \geq 1$, the growth rate of worst-case regret is bounded by $\Omega(\sqrt{T})$, and hence MSPSA achieves the optimal rate of $\Theta(\sqrt{T})$.

Theorem 7. For any value of $K \geq 1$, there exist some constants $C, C' > 0$ such that, for any policy μ ,

$$\hat{\mathcal{E}}_T^\mu \bar{\mathcal{E}}_T^\mu \geq C^2 T, \quad (2.19)$$

$$\bar{\mathcal{E}}_T^\mu \geq C\sqrt{T}, \quad (2.20)$$

and

$$\bar{\mathcal{R}}_T^\mu \geq C'\sqrt{T}. \quad (2.21)$$

Proof. See Appendix. □

2.4 Dynamic Pricing for Demand Response

In this section, we explain how the dynamic pricing problem of an electricity retailer maps to the online learning problem of Markov jump affine model in detail. To do this, we first introduce a real-time pricing mechanism that allows the electricity retailer to set the hourly retail price one day in advance and allows consumers to adjust their consumption ahead of time. Then, we present the demand model arising from the optimization problem of consumers based on thermal load dynamics and the dynamic pricing mechanism discussed. We conclude the section by discussing that the retail surplus optimization can be formulated either as a target matching problem or as a revenue maximization problem depending on the underlying assumptions.

2.4.1 Dynamic Pricing Model

The day-ahead hourly pricing mechanism, that has been used in practice by a number of utilities [20,65], works as follows:

- The retailer posts next day's hourly retail price vector $x \in \mathbb{R}^{24}$, where i th entry corresponds to i th hour price, to its customers one day ahead and keeps it fixed.
- A consumer participating the program optimizes its real-time consumption based on x and makes a payment to the retailer in the amount of x times its real-time consumption.
- In the real-time wholesale market, the retailer purchases the amount of actual demand that deviates from the day-ahead schedule by paying the real-time wholesale price.

This mechanism provides customers 24 hour price certainty as well as service guarantee. It gives customers the freedom to optimize their consumption and reduce their electricity bills.

2.4.2 Price Responsive Demand

Thermal load represents a significant part of residential electricity consumption. Therefore, we consider a price responsive demand based on the thermal load dynamics. By solving a stochastic optimization problem with the aim of maximizing the consumer utility which is a linear combination of the cost of

electricity and the expected discomfort level, in [39], it is shown that the optimal aggregate residential demand can be expressed as an affine function of the retail price.

Since the consumer utility in the case of thermal load depends on weather conditions, we assume that the affine function parameters of the aggregate demand follow a Markov jump process with a finite state space $\mathcal{S} = \{1, \dots, K\}$ and a state transition probability matrix $P = [p_{i,j}]$. Here, the state of the day is determined by the exogenous factors (which are independent of the retail price) such as the weather condition at that day. Let $s_t \in \mathcal{S}$ and $x_t \in \mathbb{R}^{24}$ be the state and the retail price vector at day t , respectively. Then, at day t , the observed real-time demand $y_t \in \mathbb{R}^{24}$ with i th entry representing the aggregate demand in i th hour from all customers in the service area is given as

$$y_t = A_{s_t}x_t + b_{s_t} + w_t \quad (2.22)$$

where $b_{s_t} \in \mathbb{R}^{24}$ and negative definite $A_{s_t} \in \mathbb{R}^{24 \times 24}$, which captures the intertemporal dependencies, are state dependent parameters of the affine demand. The random vector $w_t \in \mathbb{R}^{24}$ captures the noise in the thermal system. Hence, the demand model corresponds to the Markov jump affine model presented in Section 2.1 where the retail price vector is the control input, and the aggregate customer demand is the observed output.

2.4.3 Online Learning for Dynamic Pricing

In a two-settlement market, the retailer commits to buy its day-ahead schedule $y_t^{\text{DA}} \in \mathbb{R}^{24}$ at the day-ahead wholesale price $\lambda_t \in \mathbb{R}^{24}$ of day t , cleared on day $t - 1$, and pays $\lambda_t^\top y_t^{\text{DA}}$. In the real-time wholesale market on day t , the retailer

purchases the amount of actual consumption of its customers that deviates from the day-ahead schedule by paying the real-time wholesale price π_t . Let $u(y)$ denote the utility of the retailer for getting served with y units of electricity in the wholesale market. Then, the retail surplus on day t can be written as $u(y_t) - \lambda_t^\top y_t^{DA} - \pi_t^\top (y_t - y_t^{DA})$.

Next, we show that the retail surplus optimization can be viewed either as a target matching problem with a quadratic cost or as a revenue maximization problem depending on the underlying assumptions. Since both of these problems are studied in the previous sections, all of the results presented previously for each objective follow.

Matching the day-ahead dispatch level

Under the assumption of a quadratic generator cost function, it has been shown that the retail surplus loss is approximately measured by the normed squared deviation of the actual demand from its day-ahead schedule, *i.e.*, $\|y_t - y_t^{DA}\|_2^2$ [39].

Consequently, for a constant day-ahead dispatch level $y_t^{DA} = y^*$, optimizing the expected T-period retail profit becomes equivalent to the optimization problem presented in 2.1 with stage cost expressed by the quadratic cost (2.3) of the regulation problem. Here, the day-ahead dispatch level corresponds to the target value for customer demand observed.

Pure profit maximization

If we set the utility $u(y_t)$ of the retailer equal to its total revenue $x_t^\top y_t$ from its customers, the retail profit (surplus) on day t can be written as $(x_t - \pi_t)^\top y_t - (\lambda_t - \pi_t)^\top y_t^{\text{DA}}$. Here, we assume that the retailer is a price taker. Hence, demand y_t is independent of the real-time wholesale price π_t given the state s_t . We also assume that the real-time wholesale price π_t has an unknown state dependent distribution $g_{s_t}(\cdot)$ with bounded mean $\bar{\pi}_{s_t}$ and bounded finite variance. Furthermore, π_t and $\pi_{t'}$ are conditionally independent given states s_t and $s_{t'}$ for any $t \neq t'$.

In this setting, the unknown system parameter vector θ includes $\{\bar{\pi}_k\}_{k=1}^K$ in addition to $\{A_k, b_k\}_{k=1}^K$, and the observation vector I_{t-1} additionally includes past day-ahead wholesale prices $\{\lambda_i\}_{i=1}^t$, real-time wholesale prices $\{\pi_i\}_{i=1}^{t-1}$, and day-ahead dispatch levels $\{y_i^{\text{DA}}\}_{i=1}^t$.

Then, the optimal price x_t^* under known model is determined by maximizing the (expected) stage profit, *i.e.*,

$$\mathcal{J}(s_{t-1}, x_t) = \sum_j p_{s_{t-1}, j} \left[(x_t - \bar{\pi}_j)^\top (A_j x_t + b_j) - (\lambda_t - \bar{\pi}_j)^\top y_t^{\text{DA}} \right],$$

and is expressed by

$$x_{s_{t-1}}^* = - \left(\sum_j p_{s_{t-1}, j} (A_j + A_j^\top) \right)^{-1} \left(\sum_j p_{s_{t-1}, j} (b_j - A_j^\top \bar{\pi}_j) \right).$$

Observe that the loss (or the gain) resulting from buying day-ahead dispatch level λ_t^{DA} from day-ahead wholesale market instead of real-time wholesale market does not affect the optimal value of the retail price here due to the risk-neutral objective* (see stage profit expression). However, the expected real-time

*Expected profit maximization objective doesn't take the risk resulting from the uncertainty

wholesale price is still an important factor to determine the optimal value of the retail price. From the expression for stage profit, one can see that if the expected real-time price is high, then it is crucial to set the retail price high as well to be able to obtain a non-negative profit. Consequently, high retail price will lead to a reduced demand.

It is easy to show that, for this setting, the stage regret and T-period regret is equivalent to the ones given in Section 2.3. The MSPSA policy can be implemented as given in Figure 2.3 with a slight change: Since the objective is to optimize profit rather than revenue, one needs to assign d_{i,t_i}^+ (line 11) and d_{i,t_i}^- (line 15) to $-(x_t - \pi_t)^\top y_t$ instead of $-x_t^\top y_t$. All of the theoretical results presented in Section 2.3 remain valid.

2.5 Simulation

We present simulation results to illustrate the growth rate of regret and the optimal-input estimate convergence of MSPSA policy both for quadratic regulation and revenue maximization problems. Note that, by these simulation examples, we can only exhibit the performance of "typical" parameters and not the worst-case performance as studied in the theoretical characterization of regret.

For a benchmark comparison, we consider the greedy least square estimate (LSE) method proposed by Anderson and Taylor [4]. At each period, the greedy LSE determines the input by using the least square estimates of system parameters as if they were the true parameters and projects it onto the set Π . In order to

in the real-time wholesale price into account.

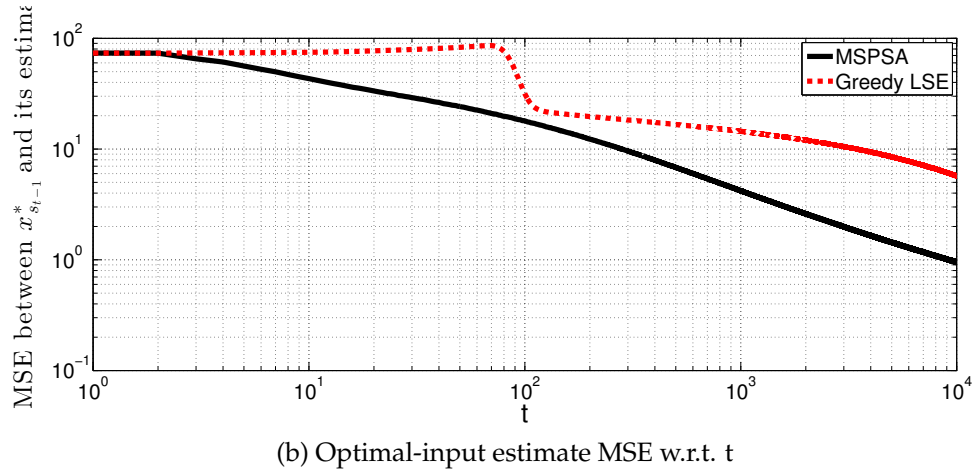
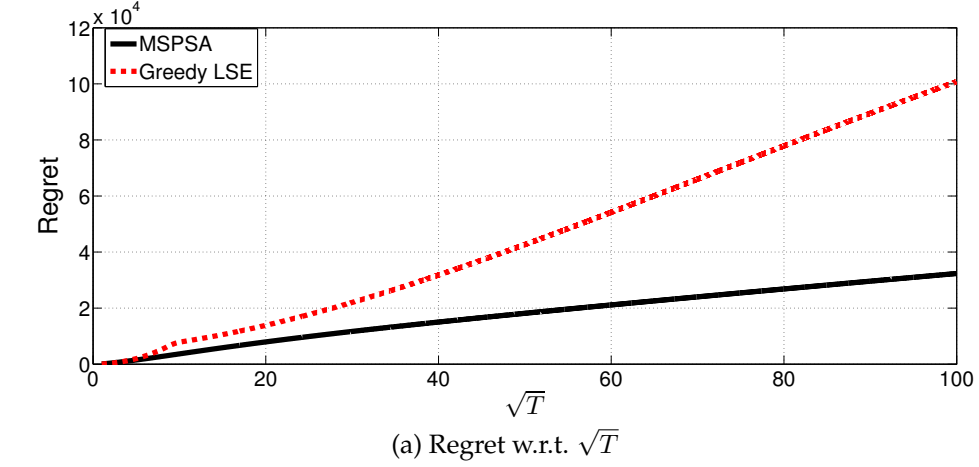


Figure 2.4: Average performance of MSPSA and Greedy LSE for quadratic regulation example.

calculate the initial LSEs of the system parameters, the first samples corresponding to the inputs generated by perturbing the initial input 5% in each direction are taken until the LSE of θ is computationally tractable. Although, in general, greedy LSE performs well numerically [4], it was shown that it can lead to incomplete learning and may not converge with positive probability which causes linear growth in regret [54].

2.5.1 Numerical Example for Quadratic Regulation Problem

To illustrate the performance of MSPSA policy for quadratic regulation problem, we consider the problem studied in [39], *i.e.*, the problem of an electricity retailer who wants to set hourly electricity prices for the next day to meet its predetermined quantity y^* in demand for each hour. Therefore, the system dimension was set to be $m = n = 24$ where each dimension corresponds to an hour of the day. Different than [39] where the demand is time-invariant, it is assumed that the demand of its customers changes depending on the state of the day, *e.g.*, weather conditions, which follows a Markov jump process. For this example, we considered two states and set the transition probability from any state to the same state to be 0.6 and to the other state to be 0.4. To calculate the average performance, 10^4 Monte Carlo runs were used.

Figure 2.4 shows the average performance of MSPSA and greedy LSE for this quadratic regulation example. The set $\Pi = [1, 4]^{24}$, and initial input was set to be the vector of all 4s for both policies. Target value y^* was taken as the vector of all 5s. For each $i \in \mathcal{S}$, A_i and b_i were chosen such that all eigenvalues of A_i belong to the interval $[-1.5, -0.5]$ and the optimal solution associated with each state is contained in Π . The noise w_t was taken as i.i.d. normal with covariance $0.5^2 I_{24}$. The MSPSA parameters were set to be $a_{t_i} = 1/(8 \times 0.5)/(t_i + 10)$ and $c_{t_i} = 1/t_i^{0.25}$; $\Delta_{t_i,j}$'s were drawn from Bernoulli(0.5) with values $\{+1, -1\}$.

In Figure 2.4a, we plot the regret of both policies with respect to square root of the time horizon. We observe that the T-period regret of MSPSA grows linearly with \sqrt{T} , which is consistent with the theoretical upper bound. On the other hand, the regret of greedy LSE seems to grow faster than linear. Therefore, we observe that MSPSA outperforms greedy LSE and the difference between the

performance of two policies is getting bigger as T increases.

Figure 2.4b illustrates how averaged normed squared error between the optimal input and its estimate changes with time in a log-log plot. From Theorem 2, we know that the optimal-input estimate and thus the input itself converges in mean square for MSPSA. In Figure 2.4b, we observe that the convergence of MSPSA is consistent with this result. Furthermore, the logarithm of the estimation error seems to decrease almost linearly with the logarithm of the time horizon. In other words, MSE seems to converge with a rate equal to $1/\sqrt{t}$. This is reasonable because in Lemma 1, we show that, for each $i \in \mathcal{S}$, the estimation error decreases with a rate equal or faster than the inverse of the square root of the number of times state i is observed. On the other hand, convergence trend for greedy LSE seems to be much slower and it performs poorly compared with MSPSA's performance.

2.5.2 Numerical Example for Revenue Maximization Problem

Here, we present an example for revenue maximization problem with system size $n = 10$ and 3 different states where the transition probability from any state to the same state was set to be 0.4 and to any other state to be 0.3. The set $\Pi = [0.75, 2]^{10}$, and initial input was set to be the vector of all 1.375s for both policies. For each $i \in \mathcal{S}$, A_i and b_i were chosen such that all eigenvalues of A_i belong to the interval $[-1.3, -0.3]$ and the optimal solution associated with each state is contained in Π . The noise w_t was taken as i.i.d. normal with covariance $0.3^2 I_{24}$. The MSPSA parameters were set to be $a_{t_i} = 1/(8 \times 0.3)/(t_i + 10)$ and $c_{t_i} = 0.75/t_i^{0.25}$; $\Delta_{t_i,j}$'s were drawn from Bernoulli(0.5) with values $\{+1, -1\}$. We

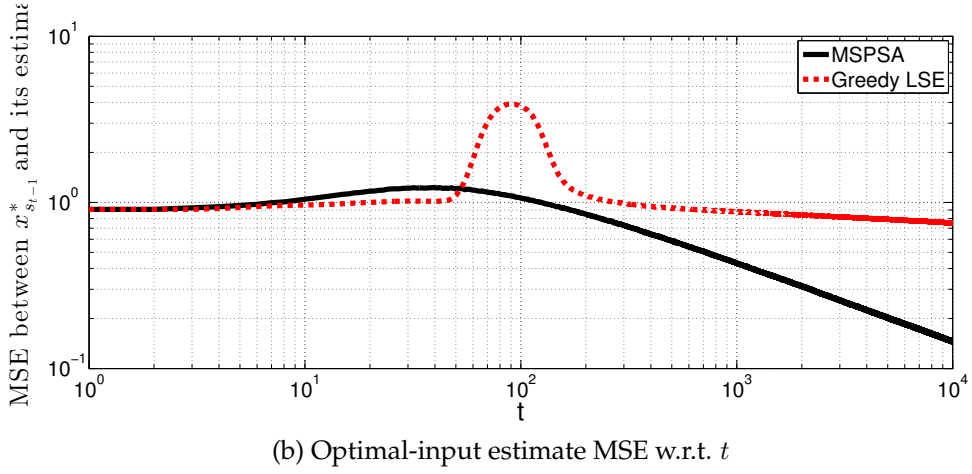
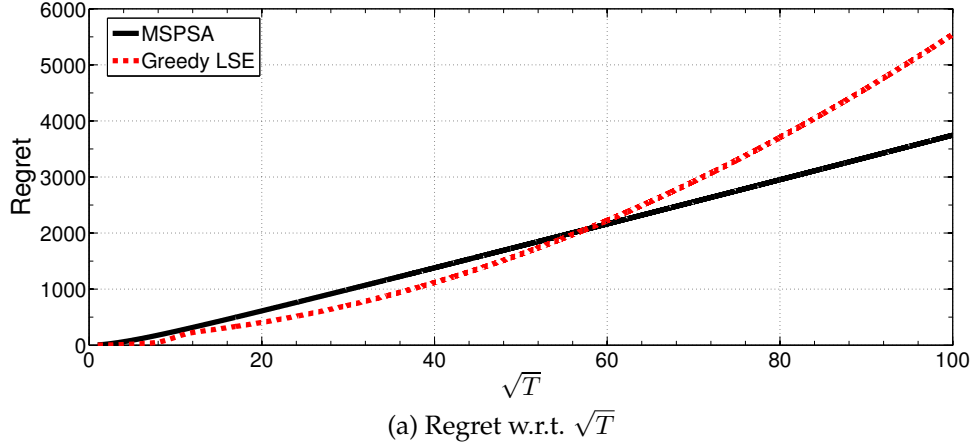


Figure 2.5: Average performance of MSPSA and Greedy ILS for revenue maximization example.

used 10^4 Monte Carlo runs to calculate the average performance of both policies.

The average performance of MSPSA and greedy LSE for this example with revenue maximization objective is given in Figure 2.5. The regret growth and the convergence of averaged normed squared error between the optimal input and its estimate are illustrated in Figure 2.5a and Figure 2.5b, respectively. In both plots, we observe a trend similar to the previous example even though the regret characterizations and the optimal inputs are different due to different objectives. Figure 2.5a shows that MSPSA's regret grows linearly with \sqrt{T}

whereas greedy LSE's regret grows almost exponentially with \sqrt{T} . Therefore, we observe that MSPSA eventually outperforms greedy LSE as T increases even though greedy LSE performs better at the beginning of the time horizon. Figure 2.5b shows that, after a slight increase at the beginning of the time horizon, the logarithm of the estimation error of MSPSA decreases linearly with the logarithm of the time horizon and becomes sufficiently small; whereas the estimation error of Greedy LSE stays almost constant except the spike that is probably due to the poor initial LSEs. Overall, we can say that MSPSA outperforms greedy LSE in both numerical examples.

CHAPTER 3

ONLINE LEARNING OF OPTIMAL BIDDING STRATEGY IN REPEATED MULTI-COMMODITY AUCTIONS

In this chapter, we develop an online learning algorithm to the problem of optimal bidding in repeated multi-commodity auctions. First, we start explaining our motivating application, virtual trading in electricity markets, in detail and present the mathematical problem formulation. Then, we explain our online learning approach and show the optimality of its regret growth rate. We conclude the chapter with a comprehensive empirical study that is conducted with the data obtained from NYISO and PJM energy markets to illustrate the performance of the proposed algorithm for the motivating example and a simulation study to illustrate the regret performance.

3.1 Virtual Trading in Electricity Markets

3.1.1 Virtual Transactions in Two-Settlement Market System

A virtual transaction on any given day (session) involves transactions in the day-ahead (DA) and real-time (RT) markets for power at a particular location and in a particular hour. Herein we refer to each location-hour pair with which a transaction is associated as a trading option. Typically, two types of virtual transactions are allowed in the US wholesale electricity markets: (i) virtual demand bid and (ii) virtual supply bid. A virtual demand bid is a bid to buy energy in the DA market with an obligation to sell back exactly the same amount

in the RT market. A virtual supply bid is a bid to sell energy in the DA market with an obligation to buy back exactly the same amount in the RT market.

The DA market takes place one day ahead of the actual power delivery. In the DA market, the independent system operator receives bids from (actual) generators and load serving entities as well as virtual bidders. After the DA market closes on day $t - 1$, the bids in the DA market are processed by the independent system operator via a security constrained economic dispatch that accepts a subset of virtual bids and determines the amount of power to generate for each generator and the associated DA prices.

The RT market takes place at the time of actual power delivery on day t . The independent system operator adjusts the dispatch level according to the actual system operating conditions and compute the RT prices. The virtual bids that are accepted in the DA market are settled in the RT market, and a virtual bidder with an accepted bid is paid at the difference of the DA and RT prices. See the more precise mathematical description of the settlement process in Sec. 3.1.2.

3.1.2 A Mathematical Model of Virtual Trading

Here we introduce a model for virtual trading. Recall that a trading option is defined by a pair of a location and a particular time of power delivery. A location can be a bus of the transmission grid or a trading zone. The time of power delivery is a specific hour in a 24 hour trading horizon.

Let $\lambda_{t,k}$ and $\pi_{t,k}$ be the DA and RT prices (in \$/MWh) of option k on day t , respectively. Let $x_{t,k}$ be a virtual bid (in \$/MWh) for option k on day t . A *virtual*

demand bid is a bid to buy a unit quantity of electricity at a particular location and hour in the DA market with the obligation to sell the same amount at the same location and hour in the RT market. The demand bid $x_{t,k}$ is cleared if the bid price $x_{t,k}$ is higher than or equal to the DA price $\lambda_{t,k}$, *i.e.*, $x_{t,k} \geq \lambda_{t,k}$. For the accepted bid, the payoff is the difference between the RT and DA prices of that option, *i.e.*,

$$(\pi_{t,k} - \lambda_{t,k})\mathbb{1}\{x_{t,k} \geq \lambda_{t,k}\}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function that is one if its argument is true and zero otherwise.

Similarly, a *virtual supply bid* is an offer to sell electricity in the DA market with the obligation to buy back in the RT market. The supply bid $x_{t,k}$ is cleared if the bid price $x_{t,k}$ is lower than or equal to the market clearing price $\lambda_{t,k}$, *i.e.*, $x_{t,k} \leq \lambda_{t,k}$. For the accepted bid, the payoff is given by

$$(\lambda_{t,k} - \pi_{t,k})\mathbb{1}\{x_{t,k} \leq \lambda_{t,k}\}.$$

The payoff for the two types of bids can be expressed by a single expression through a simple translation. To this end, we assume that DA prices are bounded with known upper/lower bounds, *i.e.* $l_\lambda < \lambda_{t,k} < u_\lambda$. Then, regardless of the type of bids, the payoff obtained from option k on day t can be written as:

$$(\pi'_{t,k} - \lambda'_{t,k})\mathbb{1}\{x'_{t,k} \geq \lambda'_{t,k}\}$$

where, for a virtual demand bid, $x'_{t,k} = x_{t,k} - l_\lambda$, $\lambda'_{t,k} = \lambda_{t,k} - l_\lambda$, and $\pi'_{t,k} = \pi_{t,k} - l_\lambda$; and, for a virtual supply bid, $x'_{t,k} = u_\lambda - x_{t,k}$, $\lambda'_{t,k} = u_\lambda - \lambda_{t,k}$, and $\pi'_{t,k} = u_\lambda - \pi_{t,k}$. In this case, observe that $x'_{t,k} = 0$ is equivalent to not bidding for option k on day t .

For notational convenience, hereafter we use $\lambda_{t,k}$, $\pi_{t,k}$, and $x_{t,k}$ instead of $\lambda'_{t,k}$, $\pi'_{t,k}$, and $x'_{t,k}$ to represent the translated price and bid variables. The accumulative return for a T-period trading horizon for a given bid $x_{t,k}$ sequence and DA/RT prices, irrespective the type of bids, is given by

$$\sum_{t=1}^T (\pi_{t,k} - \lambda_{t,k}) \mathbb{1}\{x_{t,k} \geq \lambda_{t,k}\}. \quad (3.1)$$

Next, we study the problem of a virtual trader who considers to bid on K options and aims to determine the optimal value of $x_{t,k}$ for each $k \in \{1, \dots, K\}$ under a budget constraint. Note that a trader can submit multiple bids for the same option, including demand and supply bids, simultaneously.

3.2 Problem Formulation

Let $\lambda_t = [\lambda_{t,1}, \dots, \lambda_{t,K}]^\top$ and $\pi_t = [\pi_{t,1}, \dots, \pi_{t,K}]^\top$ be the vector of auction clearing (day-ahead) and spot (real-time) prices at period (day) t , respectively. Similarly, let $x_t = [x_{t,1}, \dots, x_{t,K}]^\top$ be the vector of bids for period t . At the end of each period, the auction clearing and spot prices of all options are observed. Therefore, before choosing the bid for period t , all the information the bidder (virtual trader) has is a vector I_{t-1} containing his observation and decision history $\{x_i, \lambda_i, \pi_i\}_{i=1}^{t-1}$.^{*} Consequently, a bidding policy μ is defined as a sequence of decision rules, *i.e.*, $\mu = (\mu_0, \mu_1, \dots, \mu_{T-1})$, such that, at period $t-1$, μ_{t-1} maps the information history I_{t-1} to the bid x_t of period t .

^{*}In virtual trading, the bid for day t needs to be chosen before observing the full vector of RT prices of day $t-1$. However, in that case, $I_{t-1} = \{x_i, \lambda_i, \pi_i\}_{i=1}^{t-2}$ can be used instead without loss of generality.

The objective is to determine a bidding policy μ that maximizes the expected cumulative payoff over T periods subject to a budget constraint for each individual period. From (3.1), the optimization problem can be written as

$$\begin{aligned} & \underset{\mu}{\text{maximize}} && \mathbb{E} \left(\sum_{t=1}^T (\pi_t - \lambda_t)^\top \mathbb{1}\{x_t^\mu \geq \lambda_t\} \right) \\ & \text{subject to} && \|x_t^\mu\|_1 \leq B, \quad \forall t = 1, \dots, T, \\ & && x_t^\mu \geq 0, \quad \forall t = 1, \dots, T, \end{aligned} \tag{3.2}$$

where x_t^μ denotes the (translated) bid determined by policy μ , $\mathbb{1}\{x_t^\mu \geq \lambda_t\}$ is the vector of indicator functions with the k th entry corresponding to $\mathbb{1}\{x_{t,k}^\mu \geq \lambda_{t,k}\}$, and B is the auction budget* of the bidder. The expectation is taken with respect to randomness in $\{\pi_t, \lambda_t\}_{t=1}^T$ and the policy μ .

The joint distribution of the auction clearing and spot prices is unknown to the bidder. Hence, it is not possible to solve the optimization problem analytically. Instead, bidder uses his observation history to obtain the optimal bid.

3.3 Online Learning Approach to Virtual Trading

In this section, we develop an algorithmic bidding strategy aimed at maximizing expected payoff by allocating a fixed budget among K options without assuming the knowledge of underlying joint distribution of the auction clearing and spot prices.

*This budget provides an upper bound to DA market spending in the case of virtual demand bids only and non-negative DA prices. However, it becomes artificial with the inclusion of virtual supply bids. In the general setting, the budget constraint restricts the number of options to bid and leads to the determination of bid values that provide the best payoff per unit of bid.

An outline of our proposed approach is in order. Since the expected payoff cannot be calculated analytically due to the unknown distribution, we consider the maximization of the sample mean payoff, which is equivalent to an empirical risk minimization (ERM) problem [79]. For fixed optimization horizon T , solving this ERM amounts to solving a multiple-choice knapsack problem [40], which is NP hard. We propose a polynomial-time approximation algorithm, referred to as dynamic programming on discrete set (DPDS). We also extend this algorithm to deal with the objective of optimizing a variant of mean-variance measure.

3.3.1 An ERM Approach

Because past auction clearing and spot prices are observable, one can calculate the (empirical) average payoff that could have been obtained up to the current period by a fixed bid $z \in \mathcal{F}$ where $\mathcal{F} = \{z \in \mathbb{R}^K : z \geq 0, \|z\|_1 \leq B\}$ is the feasible set of bids. Let z_k denote the fixed bid value for option k , *i.e.*, the k th entry of the fixed bid vector z . Specifically, the average payoff $\bar{r}_{t,k}(z_k)$ from option k with fixed bid z_k in t periods is

$$\bar{r}_{t,k}(z_k) = \frac{1}{t} \sum_{i=1}^t (\pi_{i,k} - \lambda_{i,k}) \mathbb{1}\{z_k \geq \lambda_{i,k}\}.$$

For example, at the end of first period, $\bar{r}_{1,k}(z_k) = (\pi_{1,k} - \lambda_{1,k}) \mathbb{1}\{z_k \geq \lambda_{1,k}\}$ as illustrated in Fig. 3.1a. For, $t \geq 2$, this can be expressed recursively;

$$\bar{r}_{t,k}(z_k) = \begin{cases} \frac{t-1}{t} \bar{r}_{t-1,k}(z_k) & \text{if } z_k < \lambda_{t,k}, \\ \frac{t-1}{t} \bar{r}_{t-1,k}(z_k) + \frac{1}{t} (\pi_{t,k} - \lambda_{t,k}) & \text{if } z_k \geq \lambda_{t,k}. \end{cases} \quad (3.3)$$

Since each observation introduces a new breakpoint*, and the value of average payoff function is constant between two consecutive breakpoints, we observe that $\bar{r}_{t,k}(z_k)$ is a piece-wise constant function with at most t breakpoints. Let the vector of order statistics of the observed auction clearing prices $\{\lambda_{i,k}\}_{i=1}^t$ and zero be $\lambda^{(t,k)} = [0, \lambda_{(1),k}, \dots, \lambda_{(t),k}]^\top$. Let $r^{(t,k)}$ be the associated vector of average payoffs where $r_j^{(t,k)}$, the j th entry of the vector $r^{(t,k)}$, is the average payoff $\bar{r}_{t,k}(\lambda_j^{(t,k)})$ for fixed bid $\lambda_j^{(t,k)}$, the j th entry of the vector $\lambda^{(t,k)}$. Then, $\bar{r}_{t,k}(z_k)$ can be expressed by the pair $(\lambda^{(t,k)}, r^{(t,k)})$ as shown in Fig. 3.1b.

For a vector y , let $y_{m:n} = (y_m, y_{m+1}, \dots, y_n)$ denote the sequence of entries from m to n . Initialize $\lambda^{(0,k)} = 0$ and $r^{(0,k)} = 0$ at the beginning of first period. Then, at each period $t \geq 1$, the pair $(\lambda^{(t,k)}, r^{(t,k)})$ can be updated recursively as follows:

$$\lambda^{(t,k)} = \left[\lambda_{1:i_{t,k}}^{(t-1,k)}, \lambda_{t,k}, \lambda_{i_{t,k}+1:t}^{(t-1,k)} \right]^\top, \quad (3.4)$$

and

$$r^{(t,k)} = \left[\frac{t-1}{t} r_{1:i_{t,k}}^{(t-1,k)}, \frac{t-1}{t} r_{i_{t,k}:t}^{(t-1,k)} + \frac{1}{t} (\pi_{t,k} - \lambda_{t,k}) \right]^\top \quad (3.5)$$

where $i_{t,k} = \max_{i: \lambda_i^{(t-1,k)} < \lambda_{t,k}} i$.

Consequently, the overall average payoff function $\bar{r}_t(z)$ is the sum of average payoff functions of individual options, *i.e.*, $\sum_{k=1}^K \bar{r}_{t,k}(z_k)$. To determine the bid for period $t+1$, let's consider the maximization of the overall average payoff function, which corresponds to the ERM approach, *i.e.*,

$$\max_{x_{t+1} \in \mathcal{F}} \bar{r}_t(x_{t+1}) = \max_{x_{t+1} \in \mathcal{F}} \sum_{k=1}^K \bar{r}_{t,k}(x_{t+1,k}). \quad (3.6)$$

Due to the piece-wise constant structure, choosing $x_{t+1,k} = \lambda_i^{(t,k)}$ for some i contributes the same amount to the overall payoff as choosing any $x_{t+1,k} \in$

*Without loss of generality, we assume that the same auction clearing price value does not occur multiple times.

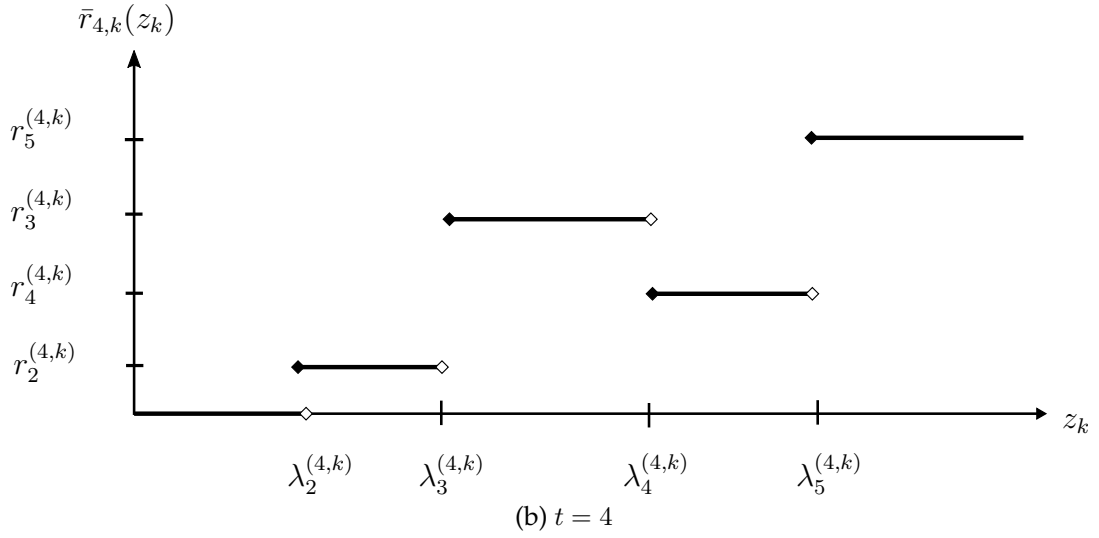
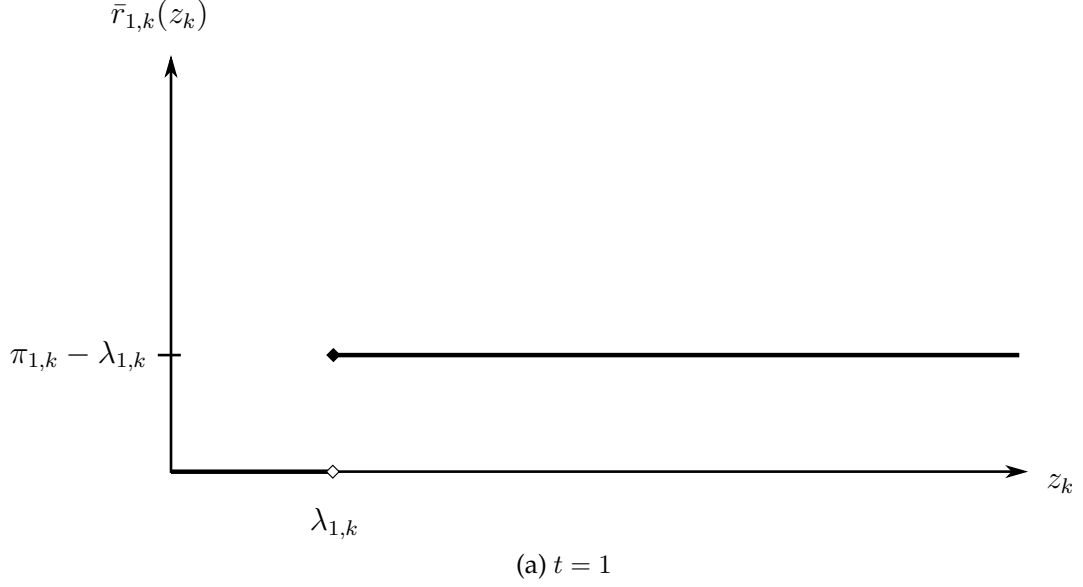


Figure 3.1: Example of a piece-wise constant average payoff function of option k

$\left[\lambda_i^{(t,k)}, \lambda_{i+1}^{(t,k)}\right)$. However, choosing $x_{t+1,k} = \lambda_i^{(t,k)}$ utilizes a smaller portion of the budget. Hence, an optimal solution to (3.6) can be obtained by solving the

following integer linear program:

$$\begin{aligned}
& \underset{\{y_k\}_{k=1}^K}{\text{maximize}} && \sum_{k=1}^K \left(r^{(t,k)}\right)^\top y_k \\
& \text{subject to} && \sum_{k=1}^K \left(\lambda^{(t,k)}\right)^\top y_k \leq B, \\
& && \|y_k\|_1 \leq 1, \quad \forall k, \\
& && y_{k,i} \in \{0, 1\}, \quad \forall (k, i).
\end{aligned} \tag{3.7}$$

where the bid value $x_{t+1,k} = \left(\lambda^{(t,k)}\right)^\top y_k$ for node k .

Observe that (3.7) is a multiple choice knapsack problem (MCKP), a generalization of the 0-1 knapsack. The MCKP problem in (3.7), unfortunately, is NP-hard [40]. Had we a polynomial-time algorithm that finds an optimal solution $x_{t+1} \in \mathcal{F}$ to (3.6), we would have obtained the solution of (3.7) in polynomial-time by setting $y_{k,i} = 1$ where $i = \max_{i: \lambda_i^{(t,k)} \leq x_{t+1,k}} i$ for each k . By contradiction, the ERM problem (3.6) is also NP-hard.

3.3.2 DPDS: A Polynomial-Time Online Learning Algorithm

We now derive a polynomial-time algorithm referred to as dynamic programming on discrete set (DPDS). The idea behind DPDS is to discretize the feasible set using intervals of equal length and optimize the average payoff on this new discrete set via a dynamic program.

Let α_t be an integer sequence increasing with t , and $\mathcal{D}_t = \{0, B/\alpha_t, 2B/\alpha_t, \dots, B\}$ is a sequence of equally placed grid points in $[0, B]$ with increasing density with t as illustrated in Fig. 3.2. Then, the new discrete set is given as $\mathcal{F}_t = \{z \in \mathcal{F} : z_k \in \mathcal{D}_t, \forall k \in \{1, \dots, K\}\}$. Our goal is to optimize $\bar{r}_t(\cdot)$ on the new set \mathcal{F}_t rather

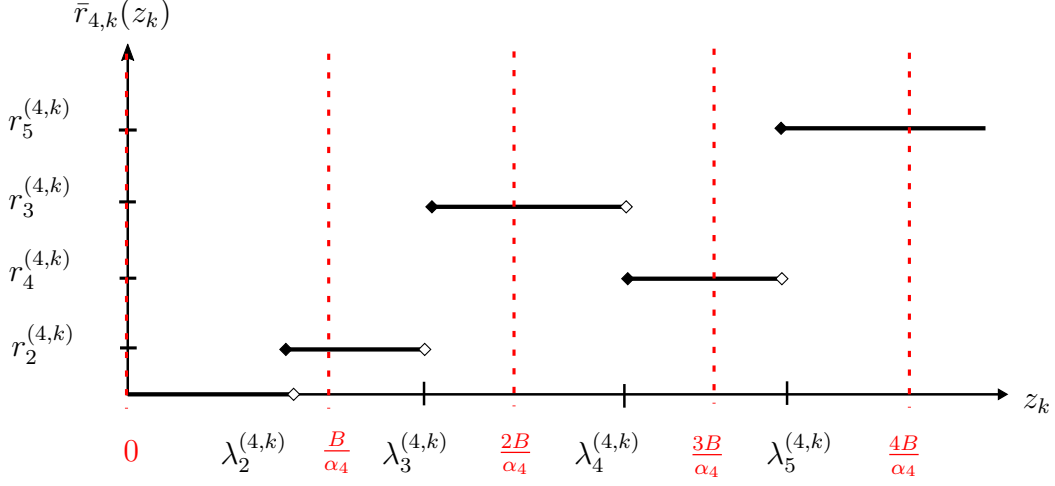


Figure 3.2: Example of the discretization of the decision space for option k when $t = 4$

than \mathcal{F} , i.e.,

$$\max_{x_{t+1} \in \mathcal{F}_t} \bar{r}_t(x_{t+1}) = \max_{x_{t+1} \in \mathcal{F}_t} \sum_{k=1}^K \bar{r}_{t,k}(x_{t+1,k}). \quad (3.8)$$

Observe that, for fixed t , this can be written as a multistage decision problem with K stages as follows: the state of stage k is the remaining budget $b_k \in D_t$, and the action (decision) of stage k is the bid value $x_{t+1,k} \in A_{t,b_k}$ of option k where $A_{t,b_k} = \{z_k \in D_t : z_k \leq b_k\}$. In this case, D_t is the state space, A_{t,b_k} the action space of stage k , and $\bar{r}_{t,k}(x_{t+1,k})$ the payoff (reward) of stage k for taking action $x_{t+1,k}$.

Now, we define the maximum payoff one can collect in state b over the remaining n stages as $V_n(b)$. Then, the Bellman equation can be used to solve for $V_K(B)$ which gives the optimal solution to (3.8). This type of dynamic programming approach has been used to solve 0-1 Knapsack problems including MCKP [32]. However, direct implementation of that approach results in pseudo-polynomial computational complexity in the case of 0-1 Knapsack problems. The discretization of the feasible set with equal interval length reduces the

computational complexity to polynomial time.

Assuming that $V_0(b) = 0$ for any b , the Bellman equation can be written as

$$V_{K-k+1}(b) = \max_{x_{t+1,k} \in A_{t,b}} (\bar{r}_{t,k}(x_{t+1,k}) + V_{K-k}(b - x_{t+1,k})), \quad (3.9)$$

which can be solved via backward induction starting from $k = K$ and proceeding toward $k = 1$. For each k , $V_{K-k+1}(b)$ is calculated for all $b \in \mathcal{D}_t$. Since the computation of $V_{K-k+1}(b)$ requires at most $\alpha_t + 1$ comparison for any fixed value of $k \in \{1, \dots, K\}$ and $b \in \mathcal{D}_t$, it has a computational complexity on the order of $K\alpha_t^2$ given the average payoff values $\bar{r}_{t,k}(x_{t+1,k})$ for all $x_{t+1,k} \in \mathcal{D}_t$ and $k \in \{1, \dots, K\}$. For each $k \in \{1, \dots, K\}$, computation of $\bar{r}_{t,k}(x_{t+1,k})$ for all $x_{t+1,k} \in \mathcal{D}_t$ introduces an additional computational complexity of at most on the order of t which can be achieved by updating $(\lambda^{(t,k)}, r^{(t,k)})$ from $(\lambda^{(t-1,k)}, r^{(t-1,k)})$ recursively as given in (3.4) and (3.5). Hence, total computational complexity of DPDS is $O(K \max(t, \alpha_t^2))$ at each period t .

3.3.3 Risk-Averse Learning

Maximizing expected profit is not necessarily a prudent strategy in algorithmic bidding. Often the risk of a particular strategy needs to be taken into account. A commonly used metric to measure the effectiveness of a strategy is the Sharpe ratio, which is the ratio of the expected return and the standard deviation of the return. In essence, this requires a trade-off between maximizing the expected return and minimizing the variance.

In order to distribute the budget among the options with high payoff and low variance, we extend DPDS algorithm to the optimization of sum of sample mean-variance of all options (a variant of the well known mean-variance

strategy [60]). The sample mean-variance function for option k can be written as

$$\begin{aligned}
\bar{r}_{t,k}^{(\rho)}(z_k) &= \bar{r}_{t,k}(z_k) - \frac{\rho}{t-1} \sum_{i=1}^t ((\pi_{i,k} - \lambda_{i,k}) \mathbb{1}\{z_k \geq \lambda_{i,k}\} - \bar{r}_{t,k}(z_k))^2 \\
&= \bar{r}_{t,k}(z_k) + \rho \frac{t}{t-1} \bar{r}_{t,k}(z_k)^2 \\
&\quad - \rho \frac{t}{t-1} \left(\frac{1}{t} \sum_{i=1}^t (\pi_{i,k} - \lambda_{i,k})^2 \mathbb{1}\{z_k \geq \lambda_{i,k}\} \right). \tag{3.10}
\end{aligned}$$

In the last equality, observe that $\bar{r}_{t,k}^{(\rho)}(z_k)$ is also a piece-wise constant function with the same breakpoints as $\bar{r}_{t,k}(z_k)$. So, the extension of DPDS follows from using $\bar{r}_{t,k}^{(\rho)}(z_k)$ instead of $\bar{r}_{t,k}(z_k)$ while solving the Bellman equation. The value of $\bar{r}_{t,k}^{(\rho)}(z_k)$ can be obtained by additionally updating the value of the last term in (3.10). This update is similar to $\bar{r}_{t,k}(z_k)$ update given in (3.5). Please see the algorithm pseudo-code given in Fig. 3.3 for the full implementation of DPDS.

3.4 Order Optimality of DPDS

We present a performance analysis of DPDS in this section. Our results are of two types. First is to show that the expected payoff of DPDS converges to the expected payoff of the globally optimal bidding strategy as the trading horizon $T \rightarrow \infty$. The second is to show that the rate of convergence of DPDS is order-optimal up to a $\sqrt{\log(T)}$ factor. This result shows that DPDS has a strong convergence property over finite optimization horizons.

Observe that $\bar{r}_{t,k}^{(\rho)}(x_{t+1,k}) = \bar{r}_{t,k}(x_{t+1,k})$ when $\rho = 0$. Here, we present our analysis for the sum of mean-variance case (for any choice of $\rho \geq 0$) that includes the expected return ($\rho = 0$) as a special case.

```

1: Initialization: Set  $x_{1,k}$ ,  $\lambda^{(0,k)}$ ,  $r^{(0,k)}$ , and  $v^{(0,k)}$  to zero  $\forall k \in \{1, \dots, K\}$ ;
2: for  $t = 1$  to  $T$  do
3:   Bid  $x_t$ ;
4:   At the end of period  $t$ , observe  $(\lambda_t, \pi_t)$ ;
5:   Update  $(\lambda^{(t,k)}, r^{(t,k)}) \forall k \in \{1, \dots, K\}$  using (3.4) and (3.5);
6:   Update  $v^{(t,k)} = \left[ \frac{t-1}{t} v_{1:i_k}^{(t-1,k)}, \frac{t-1}{t} v_{i_k:t}^{(t-1,k)} + \frac{1}{t} (\pi_{t,k} - \lambda_{t,k})^2 \right]^\top \forall k \in \{1, \dots, K\}$ 
   where  $i_k = \max_{i: \lambda_i^{(t-1,k)} < \lambda_{t,k}} i$ ;
7:   Set  $V_0(jB/\alpha_t) = 0 \forall j \in \{0, 1, \dots, \alpha_t\}$ ;
8:   Set  $V_n(0) = 0$  and  $w_n(0) = 0 \forall n \in \{1, \dots, K\}$ ;
9:   for  $k = K$  to  $1$  do
10:     $l = 2, d = 0$ , and  $j' = \alpha_t$ ;
11:    for  $j = 1$  to  $\alpha_t$  do
12:      while  $d = 0$  do
13:        if  $\lambda_l^{(t,k)} > jB/\alpha_t$  then
14:           $\bar{r}_{t,k}(jB/\alpha_t) = r_{l-1}^{(t,k)} + \rho \frac{t}{t-1} [(r_{l-1}^{(t,k)})^2 - v_{l-1}^{(t,k)}]$ ;
15:          break;
16:        else
17:          if  $l = t + 1$  then
18:             $\bar{r}_{t,k}(jB/\alpha_t) = r_l^{(t,k)} + \rho \frac{t}{t-1} [(r_l^{(t,k)})^2 - v_l^{(t,k)}]$ ;
19:             $d = 1$  and  $j' = j$ ;
20:            break;
21:          else
22:             $l = l + 1$ ;
23:          end if
24:        end if
25:      end while
26:       $V_{K-k+1}(jB/\alpha_t) = V_{K-k}(jB/\alpha_t)$ ;
27:       $w_k(jB/\alpha_t) = 0$ ;
28:      for  $i = 1$  to  $\min\{j, j'\}$  do
29:        if  $V_{K-k+1}(jB/\alpha_t) < \bar{r}_{t,k}(iB/\alpha_t) + V_{K-k}((j-i)B/\alpha_t)$  then
30:           $V_{K-k+1}(jB/\alpha_t) = \bar{r}_{t,k}(iB/\alpha_t) + V_{K-k}((j-i)B/\alpha_t)$ ;
31:           $w_k(jB/\alpha_t) = iB/\alpha_t$ ;
32:        end if
33:      end for
34:    end for
35:  end for
36:   $B_r = B$ ;
37:  for  $k = 1$  to  $K$  do
38:     $x_{t+1,k} = w_k(B_r)$ ;
39:     $B_r = B_r - x_{t+1,k}$ ;
40:  end for
41: end for

```

Figure 3.3: DPDS algorithm pseudo-code.

3.4.1 Optimal Bidding Strategy under Known Distribution and Regret

For performance analysis, it is necessary to make several assumptions. These assumptions do not limit the implementation of the algorithm; they are necessary to make the performance guarantee of DPDS precise. The assumptions, (A1), (A2), and (A3) that are used for performance analysis are given below.

Assumption (A1). *The auction clearing and spot prices (λ_t, π_t) are drawn independently* and identically† over time t from an unknown joint distribution $f(\lambda_t, \pi_t)$.*

Assumption (A2). *The payoff resulting from bidding on any option $k \in \{1, \dots, K\}$ is a bounded random variable with support in $[l, u]$ for any $z \in \mathcal{F}$, i.e. $l \leq (\pi_{t,k} - \lambda_{t,k})\mathbb{1}\{z_k \geq \lambda_{t,k}\} \leq u$.*

Define the expected payoff at period t of option k given the bid $x_{t,k}$ as

$$r_k(x_{t,k}) = \mathbb{E}((\pi_{t,k} - \lambda_{t,k})\mathbb{1}\{x_{t,k} \geq \lambda_{t,k}\} | x_{t,k}),$$

and the variance of the payoff of option k given the bid $x_{t,k}$ as

$$v_k(x_{t,k}) = \mathbb{E}(((\pi_{t,k} - \lambda_{t,k})\mathbb{1}\{x_{t,k} \geq \lambda_{t,k}\} - r_k(x_{t,k}))^2 | x_{t,k}).$$

Then, the sum of mean-variance of all options will be given by

$$r^{(\rho)}(x_t) = \sum_{k=1}^K (r_k(x_{t,k}) - \rho v_k(x_{t,k})).$$

*For a similar assumption in virtual trading literature, see Jha and Wolak [38], who showed that one cannot reject the hypothesis that the autocorrelation matrices of DA-RT price differences beyond first lag are zero. Hence, the assumption is reasonable due to prices of day $t - 1$ being unobservable before bidding for day t in reality.

†This implies that the auction clearing price is independent of x_t , which is reasonable for any market where an individual has negligible impact on the market price.

Assumption (A3). $r^{(\rho)}(.)$ is Lipschitz continuous on \mathcal{F} with p -norm and Lipschitz constant L .

Observe that if auction clearing and spot prices have a bounded support and the distribution $f(\lambda_t, \pi_t)$ is uniformly continuous and uniformly bounded on the union of that support and the feasible set \mathcal{F} , then assumptions (A2) and (A3) are satisfied.

For $\rho \geq 0$, the problem of the bidder is to find a bidding policy μ such that

$$\max_{\mu: x_t^\mu \in \mathcal{F} \forall t} \mathbb{E} \left(\sum_{t=1}^T r^{(\rho)}(x_t^\mu) \right), \quad (3.11)$$

which is equivalent to (3.2) when $\rho = 0$. Due to (A1), optimal solution to (3.11) under known distribution of (π_t, λ_t) does not depend on t and is given by

$$x^* = \arg \max_{x \in \mathcal{F}} r^{(\rho)}(x).$$

Optimal solution x^* may not be unique or it may not have a closed form. The following example illustrates a case where there isn't a closed form solution and shows that, even in the case of known distribution, the problem is a combinatorial stochastic optimization, and it is not easy to calculate an optimal solution.

Example. Let's take $\rho = 0$. Let λ_t and π_t be independent, $\lambda_{t,k}$ be exponentially distributed with mean $\bar{\lambda}_k > 0$, and the mean of $\pi_{t,k}$ be $\bar{\pi}_k > 0$ for all $k \in \{1, \dots, K\}$. Since not bidding for good k is optimal if $\bar{\pi}_k \leq 0$, we exclude the case $\bar{\pi}_k \leq 0$ without loss of generality. For this example, we can use the concavity of $r^{(0)}(x)$ in the interval $[0, \bar{\pi}]$, where $\bar{\pi} = [\bar{\pi}_1, \dots, \bar{\pi}_K]^\top$, to obtain the unique optimal solu-

tion x^* , which is characterized by

$$x_k^* = \begin{cases} \bar{\pi}_k & \text{if } \sum_{k=1}^K \bar{\pi}_k \leq B, \\ 0 & \text{if } \sum_{k=1}^K \bar{\pi}_k > B \text{ and } \bar{\pi}_k / \bar{\lambda}_k < \gamma^*, \\ z_k \text{ satisfying } (\bar{\pi}_k - z_k) e^{-z_k / \bar{\lambda}_k} / \bar{\lambda}_k = \gamma^* & \text{if } \sum_{k=1}^K \bar{\pi}_k > B \text{ and } \bar{\pi}_k / \bar{\lambda}_k \geq \gamma^*, \end{cases}$$

for all $k \in \{1, \dots, K\}$ where x_k^* denotes the k th entry of x^* , and the Lagrange multiplier $\gamma^* > 0$ is chosen such that $\|x^*\|_1 = B$ is satisfied. This solution takes the form of a "water-filling" strategy. More specifically, if the budget constraint is not binding, then the optimal solution is to bid $\bar{\pi}_k$ for every good k . However, in the case of a binding budget constraint, the optimal solution is determined by the bid value at which the marginal expected payoff associated with each good k is equal to $\min(\gamma^*, \bar{\pi}_k / \bar{\lambda}_k)$, and this bid value cannot be expressed in closed form.

Following the online machine learning literature, we measure the performance of any bidding policy μ by its regret $\mathcal{R}_T^\mu(f)$, defined by the difference between the total expected payoff of policy μ and that of the optimal solution x^* , i.e.,

$$\mathcal{R}_T^\mu(f) = \sum_{t=1}^T \mathbb{E} (r^{(\rho)}(x^*) - r^{(\rho)}(x_t^\mu)).$$

By definition, the regret is monotonically increasing for any policy μ and grows linearly with T for the worst possible μ . Since we define optimality as maximizing the expected payoff, observe that a policy μ converges to the optimal bidding strategy if the incremental regret $\mathbb{E} (r^{(\rho)}(x^*) - r^{(\rho)}(x_t^\mu))$ goes to zero as $t \rightarrow \infty$.

3.4.2 Convergence and Regret Bound for DPDS

Theorem 8 below shows that the expected payoff of DPDS converges to the expected payoff of the optimal solution x^* . More precisely, it characterizes the rate of convergence and the regret growth rate of DPDS.

Theorem 8. *Let x_{t+1}^{DPDS} denote the bid of DPDS policy for period $t + 1$. Let DPDS parameter choice $\alpha_t = \max(\lceil \alpha t^\gamma \rceil, 2)$ with $\gamma \geq 1/2$ and $\alpha > 0$, and let (A1), (A2), and (A3) hold. Then, for $t \geq 2$,*

$$\mathbb{E}(r^{(\rho)}(x^*) - r^{(\rho)}(x_{t+1}^{DPDS})) \leq C_1 \sqrt{\log t/t} + C_2 t^{-1/2}$$

and for $T > 1$,

$$\mathcal{R}_T^{DPDS}(f) \leq C \sqrt{T \log T},$$

where $C = 2(C_1 + C_2)$ and C_1 and C_2 are positive constants which depend on the values of $K, L, p, B, u, l, \rho, \alpha$, and γ .

Proof. See the appendix. □

The proof of Theorem 8 is derived by showing that the expected payoff of $x_{t+1}^* = \arg \max_{x \in \mathcal{F}_t} r^{(\rho)}(x)$ converges to that of x^* due to Lipschitz continuity, and the expected payoff of x_{t+1}^{DPDS} converges to that of x_{t+1}^* via the use of McDiarmid's inequality.

3.4.3 Lower Bound of Regret for any Bidding Policy

We now show that DPDS in fact achieves the slowest possible regret growth. Specifically, Theorem 9 states that the regret of any policy is lower bounded by

$\Omega(\sqrt{T})$. This result implies that the convergence rate of the expected payoff for any policy cannot be faster than $\Omega(1/\sqrt{t})$ because, otherwise, the regret growth would have been slower than $\Omega(\sqrt{T})$. Hence, DPDS achieves the order-optimal convergence as well as the slowest possible regret growth rate up to a logarithmic factor.

Theorem 9. *Consider the case where $K = 1$, $B = 1$, $\rho = 0$. For any bidding policy μ , there exists a distribution f satisfying assumptions (A1), (A2), and (A3) such that*

$$R_T^\mu(f) \geq \frac{1}{16\sqrt{5}}\sqrt{T}.$$

Proof. See the appendix. □

The proof of Theorem 9 is derived by showing that, every time the bid is cleared, an incremental regret greater than $T^{-1/2}/(4\sqrt{5})$ is incurred under a distribution; otherwise, the same incremental regret is incurred under another distribution. However, to distinguish between these two distributions, one needs $\Omega(T)$ samples which results in a regret lower bound given in Theorem 9. The bound is obtained by adapting a similar argument used by [8] in the context of non-stochastic multi-armed bandit problem.

3.5 Empirical Study

3.5.1 Setup and Data

For the empirical study, we consider virtual bids on zonal nodes for two different independent system operators: NYISO and PJM. We use historical DA and

RT price data from the beginning of 2006 until the end of 2016 of NYISO and PJM zones. This data set is available for all 11 zones of NYISO and for 19 zones of PJM. Since the price varies in time and location, there are $N \times 24$ different trading options every day where $N = 11$ for NYISO and $N = 19$ for PJM. The prices are per unit (MWh) prices. We consider virtual demand and virtual supply bids simultaneously for all options by using the model presented in Sec. 3.1.2. This model requires the knowledge of an upper bound u_λ and a lower bound l_λ for DA price. We choose u_λ and l_λ accordingly for each independent system operator by looking at the range of the historical DA prices in that markets. We set $u_\lambda = 1000$ and $l_\lambda = 0$ for NYISO; and $u_\lambda = 1050$ and $l_\lambda = -30$ for PJM. Consequently, total number of options is $K = 2 \times N \times 24$.

The DA market for day t closes early in the morning on day $t - 1$ for both NYISO and PJM. Hence, all of the RT prices of day $t - 1$ cannot be observed before the bid submission for day t . Therefore, the most recent observation used for any algorithm was from day $t - 2$ to determine the bid for day t .

3.5.2 Benchmark Methods

We compare DPDS with three algorithms. One is UCBID-GR inspired by UCBID [80]. On each day, UCBID-GR sorts all trading options according to their profitabilities, *i.e.*, their historical average DA-RT price spreads. Then, starting from the most profitable option, it sets the bid for an option equal to its historical average RT price* until there isn't any sufficient budget left.

*The bid is set to zero if historical average RT price is negative because bidding less than or equal to zero implies not bidding on that option.

The second algorithm is a variant of Kiefer-Wolfowitz stochastic approximation method, herein referred to as SA. SA approximates the gradient of the payoff function by using the current observation and updates the bid of each k as follows;

$$x_{t+1,k} = x_{t,k} + a_t \frac{(\pi_{t-1,k} - \lambda_{t-1,k})(\mathbb{1}\{x_{t,k} + c_t \geq \lambda_{t-1,k}\} - \mathbb{1}\{x_{t,k} - c_{t,k} \geq \lambda_{t-1,k}\})}{c_t}.$$

Then, x_{t+1} is projected to the feasible set \mathcal{F} . The step size a_t and c_t of SA were determined by searching for values that provide relatively better payoff and were set as $20000/(t-1)$ and $2000/(t-1)^{0.25}$, respectively.

The last algorithm is SVM-GR, which is inspired by the use of support vector machines (SVM) by Tang et al. [75] to determine if a demand or a supply bid is profitable for an option, *i.e.*, if the price spread is positive or negative. Due to possible correlation of a particular option's price spread on any given day with the price spreads of that and also of other options that are observed recently, for day t , the input of SVM for each option is set as the price spreads of all options from day $t-7$ to day $t-2$. To test SVM-GR algorithm at a particular year, for each option, the data from the previous year is used to train SVM and to determine the average profit, *i.e.*, average price spread, and the bid level that will be accepted with 95% confidence in the event that a demand or a supply bid is profitable. For the test year, on each day, SVM-GR first determines if a demand or a supply bid is profitable for each option. Then, SVM-GR sorts all options according to their average profits, and, starting from the most profitable option, it sets the bid of an option equal to the bid level with 95% confidence of acceptance until there isn't any sufficient budget left.

The DPDS algorithm was tested for ρ values of 0 and 0.002 to evaluate the performance under a sum of mean-variance objective instance as well as for

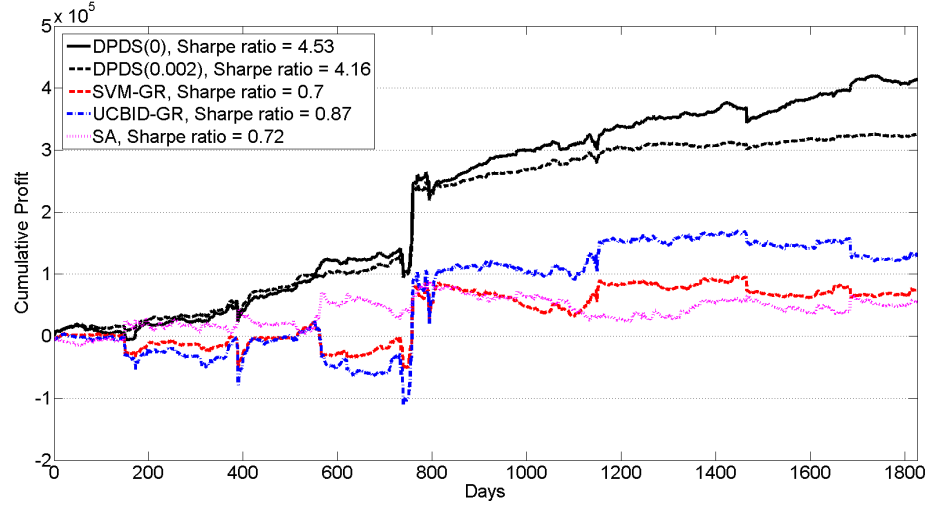
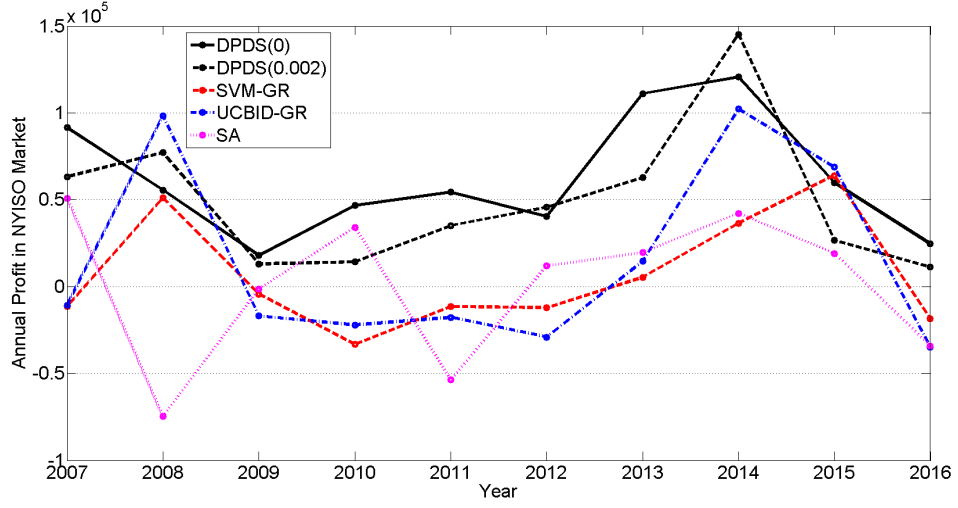


Figure 3.4: Cumulative profit trajectory from 2012 to 2016 in NYISO for $B=\$250,000$ after an initial training with 2011 data.

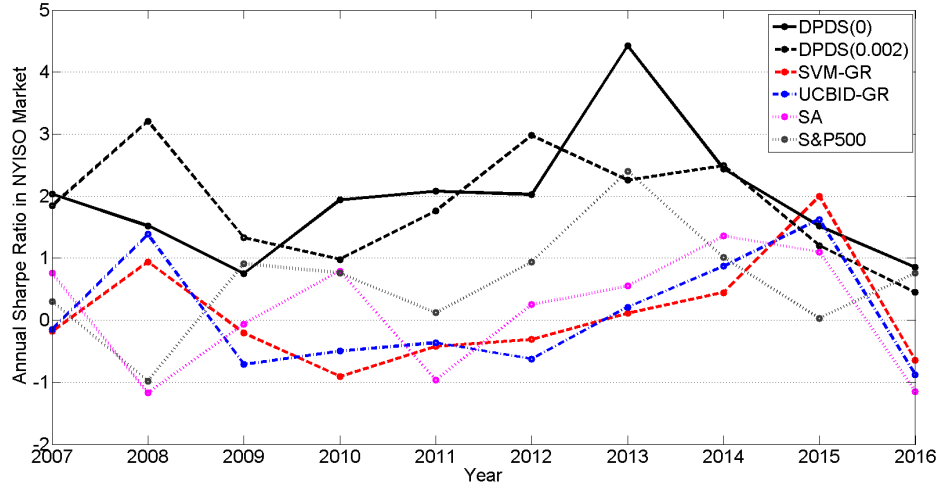
$\rho = 0$ (the risk-neutral objective). To differentiate between these two different choices of ρ , let $\text{DPDS}(\rho)$ denote the DPDS algorithm with associated ρ value. The DPDS algorithm parameter α_t was set to be $t - 1$.

3.5.3 Empirical Results

For each algorithm, the trajectory of cumulative profit that was obtained in NYISO market with a daily budget of $B=\$250,000$ from the beginning of 2012 until the end of 2016 is given in Fig. 3.4. Since the data of 2011 was required to train SVM-GR, other algorithms were also trained starting from the beginning of 2011. First, we observed that DPDS significantly outperformed other algo-



(a) Annual profit versus year



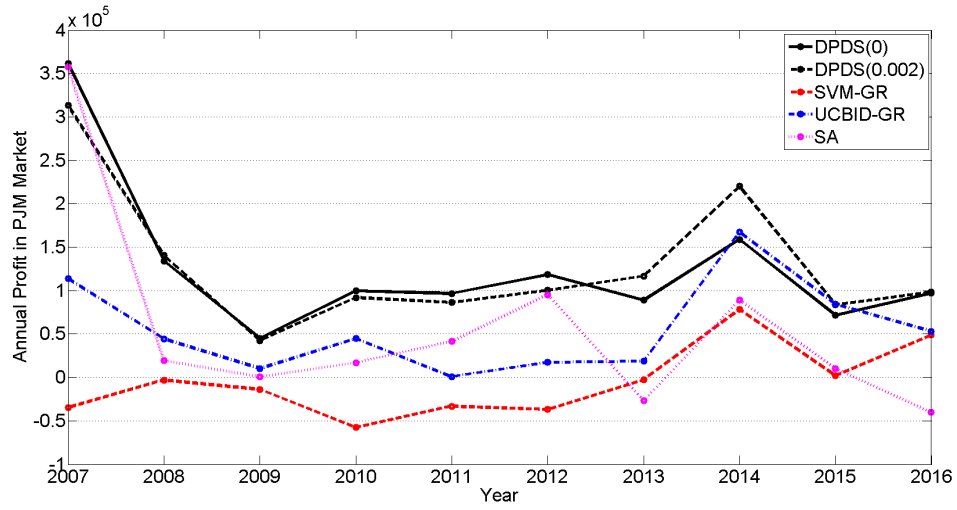
(b) Annual Sharpe ratio versus year

Figure 3.5: Annual performance in NYISO for $B = \$250,000$ (For each year, an initial training with previous year's data was performed.)

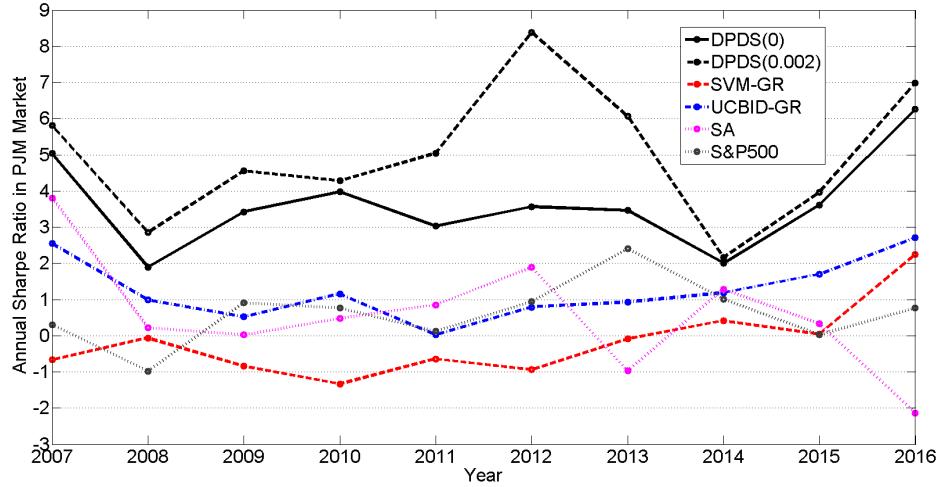
rithms in terms of Sharpe ratio*, including the S&P 500 Sharpe ratio[†] of 2.10 for

*Here, Sharpe ratio is calculated as $\sqrt{T(T-1)}\bar{r}_T / \sqrt{\sum_{t=1}^T (r_t - \bar{r}_T)^2}$ where $\bar{r}_T = \frac{1}{T} \sum_{t=1}^T r_t$, T is the number of trading days during the time period under consideration, and r_t is the percentage return of day t , which is equal to the profit of day t for virtual trading with fixed daily budget.

[†]To calculate this, S&P 500 adjusted closing price data for the time period under consideration is used. This data is obtained from Yahoo finance.



(a) Annual profit versus year

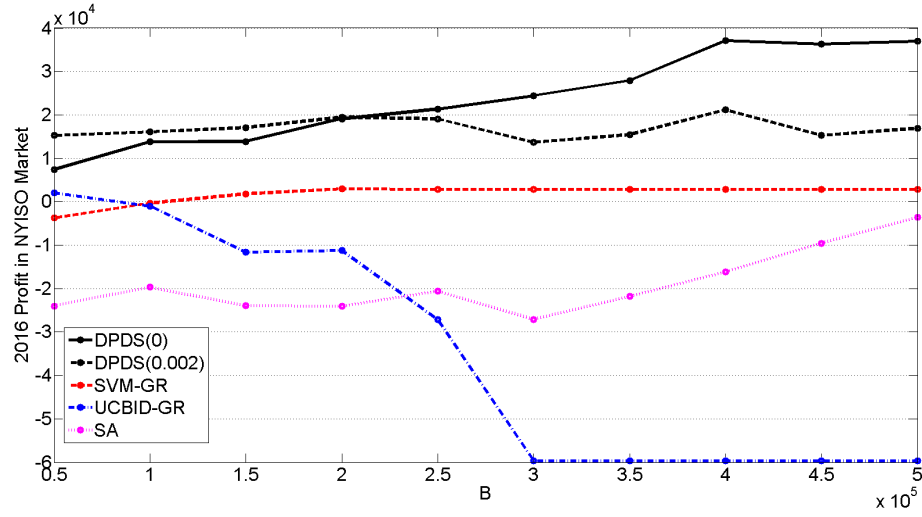


(b) Annual Sharpe ratio versus year

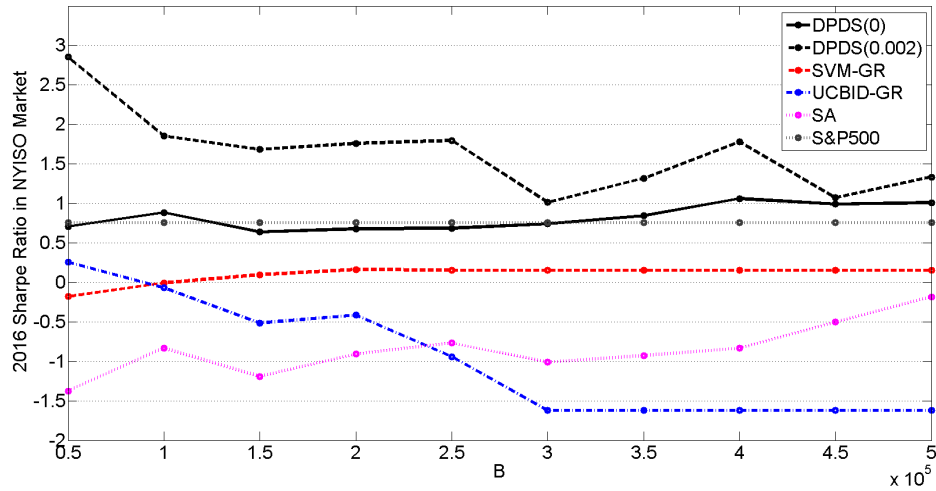
Figure 3.6: Annual performance in PJM for $B = \$250,000$ (For each year, an initial training with previous year's data was performed.)

the same period. See the legend of Fig. 3.4. This showed the significant value of participating in virtual trading in terms of profitability and risk trade-off.

The cumulative profit of DPFS, as shown in Fig. 3.4, outperformed all other algorithms with DPDS(0), which generated the highest profit. Comparing DPDS(0) and DPDS(0.002) with the latter taking into account the variance of



(a) 2016 profit versus budget level



(b) 2016 Sharpe ratio versus budget level

Figure 3.7: 2016 Performance in NYISO under different budget levels after an initial training with 2014 and 2015 data

the return, we observed from Fig. 3.4 that DPDS(0.002) generated a smoother return trajectory by avoiding more risky bids and generating less profit. We observed that, even though other algorithms were profitable; the increase in their cumulative profits was not consistent. Particularly, for UCBID-GR and SVM-GR, most of their profit resulted from a jump occurred in January 2014 due to a polar vortex [67], which didn't affect SA because of SA's incremental bid update

via a local search.

To gain insights from the performance of these algorithms on a yearly basis, annual performances for 10 consecutive years in NYISO market and in PJM market are provided in Fig. 3.5 and in Fig. 3.6, respectively. To evaluate the performance of a given year, SVM-GR used the data from the previous year for training. Hence, all other algorithms were trained for each year starting from the beginning of the previous year. Fig. 3.5(a) illustrates the total profit that is obtained each year in NYISO. We observed that DPDS outperformed all other algorithms almost every year and consistently achieved a positive profit each year for both ρ values; whereas, all other algorithms incurred losses frequently. Due to the increasing trend in profits from 2009 to 2014, we couldn't conclude that there was a decrease in profits over the years as a result of price convergence despite the decrease in the last two years. In NYISO, 2016 seemed to be the worst year in terms of profitability in general. Annual Sharpe ratios of all algorithms along with that of S&P 500 are illustrated in Fig. 3.5(b) for NYISO. We observed that DPDS outperformed other algorithms and S&P 500 also in terms of Sharpe ratio.

Similarly, total profit and Sharpe ratios that were achieved each year in the PJM market are provided in Fig. 3.6(a) and Fig. 3.6(b), respectively. In PJM, we observed that the trends in terms of both profit and Sharpe ratio were similar to the ones observed in NYISO. In general, we observed that the profit margins of all algorithms except SVM-GR were much higher in PJM compared with NYISO. Similar to the case in NYISO, in PJM, DPDS achieved higher Sharpe ratios than any other algorithm and than S&P 500. However, in PJM, the performance gap between DPDS and others was much more significant. Furthermore, the

Sharpe ratios were in general higher for all algorithms (except SVM-GR) in PJM compared with NYISO counterparts. In PJM, especially DPDS exhibited very high Sharpe ratios, *i.e.*, between 2 and 9, which were consistently higher for $\rho = 0.002$ (around 5 on average) compared with $\rho = 0$ (around 3.6 on average).

To illustrate how algorithms performed under different budget constraints, we examined the NYISO market in 2016, the year with the lowest levels of profit and Sharpe ratio (see Fig. 3.5). Total profit and Sharpe ratio for this period under different budget levels are illustrated in Fig 3.7(a) and Fig 3.7(b), respectively. Here, all algorithms were trained initially with the data from the previous two years rather than only previous year. When we increased the data used for initial training to two years, we observed that algorithms performed significantly better in terms of both profit and Sharpe ratio in general. We observed that DPDS outperformed other algorithms at all budget levels, and profit of DPDS(0) increased with increasing budget; whereas the profit of DPDS(0.002) stayed in the same range without an increasing trend. This was reasonable because DPDS(0) optimized profit and should exhibit a profit increase for higher budgets; whereas DPDS(0.002) optimized a linear combination of profit and variance term, which did not indicate a profit increase. SVM-GR also illustrated an increasing trend in profit, but this trend was much smaller compared with the trend of DPDS(0). For both SA and UCBID-GR, big losses were observed almost at all budget levels. In Fig. 3.7(b), we observed that DPDS achieved higher Sharpe ratios than other algorithms for both ρ values, and the Sharpe ratio of DPDS(0) stayed around the Sharpe ratio of S&P 500; whereas DPDS(0.002) achieved higher Sharpe ratio than DPDS(0) consistently. So, even though the profit levels of 2016 were not as high as the ones that were obtained in previous years, there were bidding strategies that achieved better Sharpe ratio than that

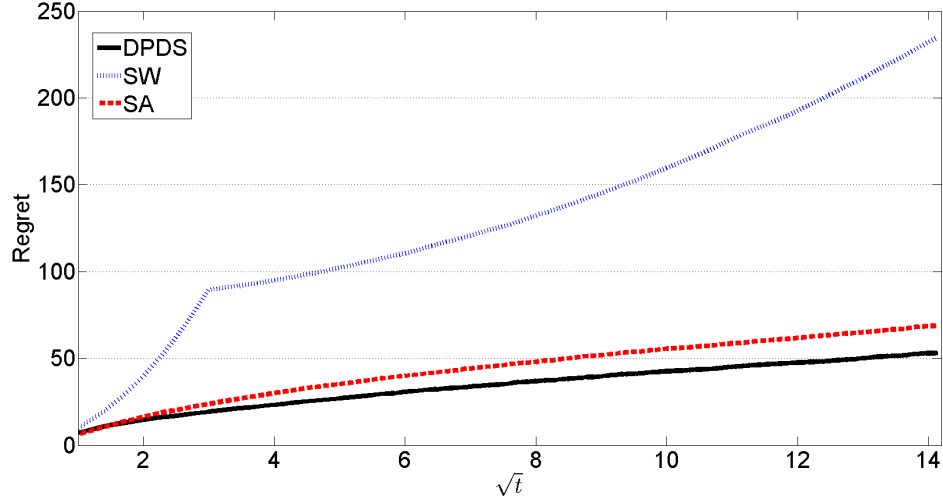


Figure 3.8: Regret with respect to \sqrt{t} when $B = 13.845$

of S&P 500.

3.6 Simulation Study

Here, we present a simulation example to illustrate the regret growth rate of DPDS. We consider an example with $K = 5$. In this example, π_t and λ_t are independent, λ_t is exponentially distributed with mean $\bar{\lambda} = [4, 6, 8, 8, 4]^\top$, and π_t is uniformly distributed with mean $\bar{\pi} = [5, 8, 8, 9, 3]^\top$ and support in $[\bar{\pi} - 1, \bar{\pi} + 1]$. Previously, in Sec. 3.4.1, we stated the characterization of the optimal solution for this example for $\rho = 0$. By using this characterization, we determined the optimal solution and the associated budget B for a range of values of the Lagrange multiplier γ^* of the budget constraint. More specifically, for the values 0.1, 0.2, 0.3, and 0.4 of γ^* , the corresponding values of B are 25.828, 20.870, 17.018, 13.845, respectively. We evaluate the performance of algorithms for these four different values of B .

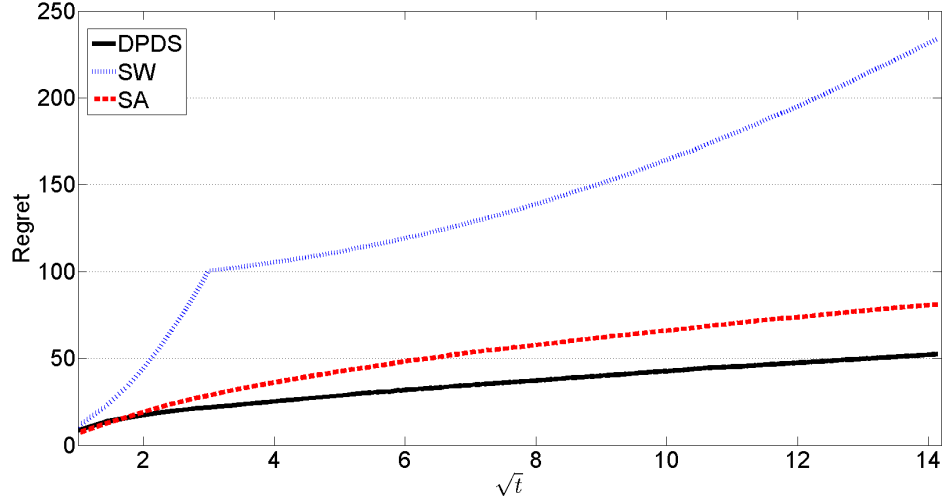


Figure 3.9: Regret with respect to \sqrt{t} when $B = 17.018$

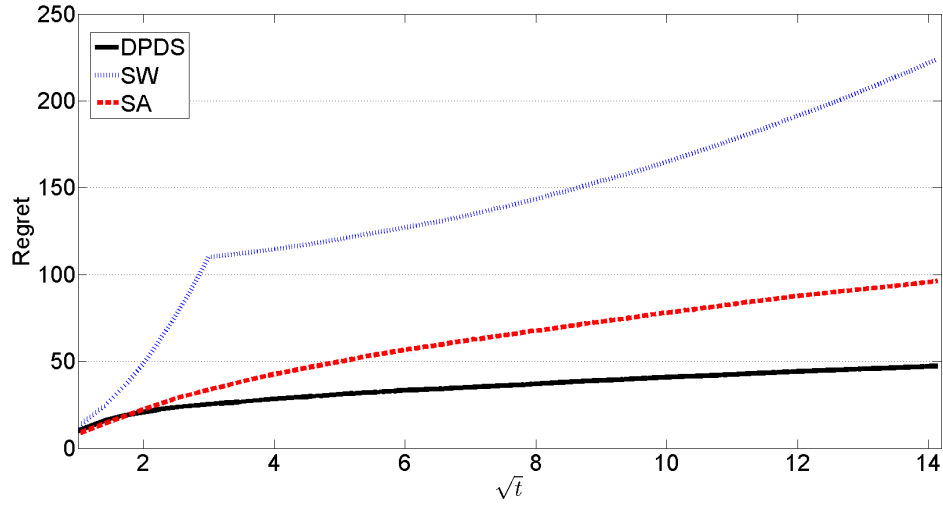


Figure 3.10: Regret with respect to \sqrt{t} when $B = 20.870$

As a benchmark comparison we consider two different approaches. The first one is based on a sliding window (SW) forecasting approach that calculates the average payoff function of each good every day from the prices of last ten days only. Then, it determines the optimal solution maximizing the total average payoff by solving the integer linear program given in (3.7). The second one, referred to as SA, is a variant of Kiefer-Wolfowitz stochastic approximation

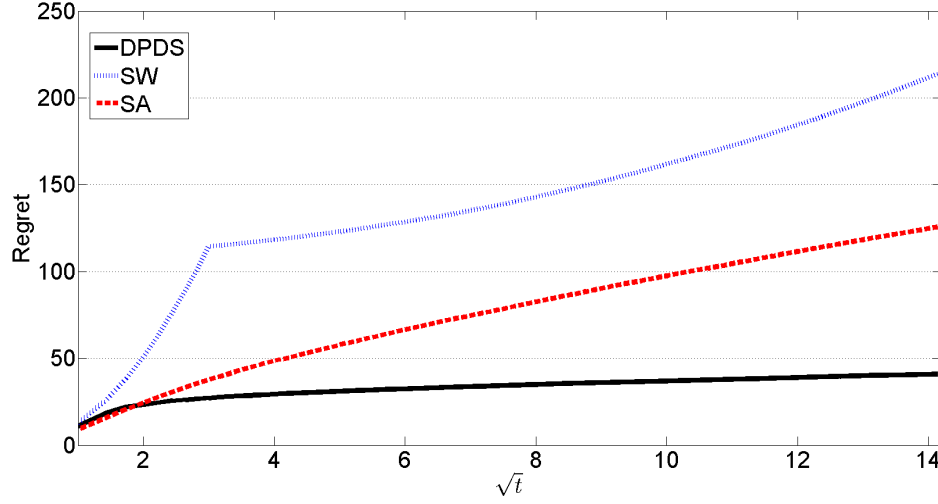


Figure 3.11: Regret with respect to \sqrt{t} when $B = 25.828$

method as explained in Sec. 3.5.2. Recall that SA approximates the gradient of the payoff function using the most recent observation and updates the bid of each k . For this example, the update step of SA is calculated as

$$x_{t+1,k} = x_{t,k} + a_t \frac{(\pi_{t,k} - \lambda_{t,k})(\mathbb{1}\{x_{t,k} + c_t \geq \lambda_{t,k}\} - \mathbb{1}\{x_{t,k} \geq \lambda_{t,k}\})}{c_t}.$$

Then, SA projects x_{t+1} to the feasible set \mathcal{F} . To give a good result for $B = 13.845$, step size a_t and c_t were carefully chosen to be $5.5/t$ and $2.5/t^{1/4}$, respectively. We set the DPDS algorithm parameter $\alpha_t = t$.

To calculate the average performance, 1000 Monte Carlo runs were used. The regret performances for budgets 13.845, 17.018, 20.870 and 25.828 are given in Fig. 3.8 through Fig. 3.11. In all cases, DPDS outperforms, and its order of regret growth is actually better than \sqrt{T} . When the SA algorithm parameters are tuned well, we observe that its performance may get close to DPDS as in Fig. 3.8. However, when we increase the budget to 25.828 gradually, the performance of SA deteriorates significantly. Also, as seen in all these figures, the regret of SW is much higher than DPDS and SA because SW does not converge to the optimal

solution due to fixed number of samples used in prediction.

CHAPTER 4

CONCLUSIONS

This thesis focuses on two online learning problems: the online learning and optimization of Markov jump affine models, and the online learning of optimal bidding strategy in repeated multi-commodity auctions. Here, we present our concluding remarks separately for each problem.

4.1 Online Learning and Optimization of Markov Jump Affine Models

In this part, we presented an online learning and optimization approach for Markov jump affine models with unknown parameters for two different objectives: (i) quadratic regulation and (ii) revenue maximization. For both objectives, we established that the MSPSA achieves the optimal rate of regret growth $\Theta(\sqrt{T})$. Compared to the classical time-invariant affine model, introducing a Markov jump process to the system does not change the optimal rate of regret growth for the revenue maximization objective. On the other hand, for the quadratic regulation objective, our result indicates a significant change on the optimal order of regret growth with the introduction of Markov jump dynamics. More specifically, the optimal regret growth rate changes from $\Theta(\log T)$ to $\Theta(\sqrt{T})$ due to the increase in the learning complexity. This result was not clear previously because it does not follow from the lower bound proof of the time-invariant setting of the problem.

The algorithms proposed in the literature for the time-invariant model are no longer applicable for the extended model with Markov jump dynamics because

the optimal input under the extended model is a function of the observed state due to the Markov chain, and the direct implementations of existing algorithms do not take into account the observed state of the Markov process.

To deal with the changing Markov process dynamics, we extend Spall's stochastic approximation method by introducing the idea of state tracking. Even though the idea of state tracking may seem to be an intuitive extension, it is not obvious if this type of extension of existing methods would actually converge to the optimal input. For example, when the same idea of state tracking in MSPSA is applied to Robbins-Monro algorithm, which is used to solve the quadratic regulation problem without Markov jump dynamics, we observe that the solution converges to a suboptimal point resulting in a linear regret growth. Yet, we show that it is possible to achieve the optimal regret growth rate with MSPSA. We also show that this policy converges to the optimal input both almost surely and in mean square. Furthermore, MSPSA has the flexibility of gradient descent type algorithms in terms of applicability to more general problems beyond the ones with linear model. However, more general classes of problems are beyond the scope of this thesis and need to be treated separately.

Besides the contributions mentioned above, derivation of the lower bound of regret provides some other interesting conclusions. One of these implications is that the regret increases from $\Theta(\log T)$ to $\Theta(\sqrt{T})$ even for the single state case of the quadratic regulation problem when the system matrix becomes non-invertible (but is still full-column). In other words, adding just one row to a square and invertible matrix changes the regret order from $\Theta(\log T)$ to $\Theta(\sqrt{T})$. This result is relevant in problems where the dimension of the observed output is greater than that of the input.

Another implication of the lower bound is an insight into the tradeoff between exploration and exploitation: the product of the cumulative estimation error and the cumulative input error grows linearly with T for any policy. This result indicates that if the goal is to optimize the estimation error only, then it is possible to find a policy for which the estimation error grows slower than \sqrt{T} in which case the input error has to grow faster than \sqrt{T} . In fact, if the MSPSA algorithm parameter c_t is set to be a constant rather than a decreasing sequence of t , by following the upper bound proof of MSPSA, it can be shown that the estimation error grows with $\log T$ whereas the input error and, hence, the regret grow linearly with T .

The assumption of Markov process being observable and exogenous holds for our motivating application: the dynamic pricing problem of an electricity retailer with consumer demand that changes according to the exogenous weather conditions. However, this assumption may become problematic for other applications. Therefore, it is also of interest to study the more general setting of hidden Markov and endogenous Markov processes. Indeed, such problems are receiving increasing attention. For example, reinforcement learning problem in Markov decision processes (MDPs) gained significant popularity due to the success of deep learning. However, the results on fundamental (theoretical) limits of learning in this area, especially, for MDPs with continuous action spaces, are extremely limited. Since our model is a special case of these more complex models including MDPs with continuous action space, the lower bound of regret in our work constitutes also a lower bound for these problems. Hence, this result provides a concrete initial step toward this general setting. Yet, it is not obvious if MSPSA policy can be utilized to achieve convergence in these more general settings. As a future work, it would be interesting to study if MSPSA policy can

be extended to deal with these settings.

4.2 Online Learning of Optimal Bidding Strategy in Repeated Multi-Commodity Auctions

We study the algorithmic bidding problem under budget constraint in repeated multi-commodity auctions. Despite the fact that the objective function involved is non-convex and the ERM problem is NP-hard, by combining general techniques such as discretization approach and dynamic programming with ERM approach, we derive a practical and efficient algorithm to the problem. We show that the expected payoff of the proposed algorithm, DPDS, converges to that of the optimal strategy by a rate no slower than $\sqrt{\log t/t}$, which results in a $O(\sqrt{T \log T})$ regret. By showing that the regret is lower bounded by $\Omega(\sqrt{T})$ for any bidding strategy, we prove that DPDS is order optimal up to a $\sqrt{\log T}$ term.

For the motivating application of virtual trading in electricity markets, the stochastic setting, studied here, is natural due to the electricity markets being competitive, which implies that the existence of an adversary is unlikely. However, it is also of interest to study the adversarial setting to extend the results to other applications. For example, the adversarial setting of our problem is a special case of no-regret learning problem of Simultaneous Second Price Auctions (SiSPA), studied in [28] and [31].

In particular, to deal with the adversarial setting, it is possible to use our dynamic programming approach as the offline oracle for the Oracle-Based Generalized FTPL algorithm proposed in [31] if we fix the discretized action set over

the whole time horizon. More specifically, let the interval length of discretization be B/m , i.e., $\alpha_t = m$. Then, it is possible to show that a 1-admissible translation matrix with $K \lceil \log m \rceil$ columns is implementable with complexity m . Consequently, no-regret result of [31] holds with a regret bound of $O(K\sqrt{T} \log m)$ if we measure the performance of the algorithm against the best action in hindsight in the discretized finite action set rather than in the original continuous action set considered here. Unfortunately, as shown by Weed et al. [80], it is not possible to achieve sublinear regret with a fixed discretization for the specific problem considered here. Hence, it requires further work to see if this method can be extended to obtain no-regret learning for the adversarial setting under the original continuous action set.

The performance of the proposed algorithm is evaluated empirically by using a large historical data set that is obtained from NYISO and PJM energy markets. Empirical results show that the proposed strategy consistently outperforms benchmark methods and achieves significant profit. More significant, perhaps, is that the proposed algorithm showed better Sharpe ratio against competitors, including the S&P 500 index. Such historical data, obviously, do not confirm with the assumption made for the regret result. This suggests a level of robustness of the proposed algorithm.

There are several directions that the proposed approach can be generalized. The algorithm presented here optimizes the bid values (willingness to pay) for options but not the quantities (number of MWhs for virtual trading). Even though the problem formulation allows optimization of multiple copies of the same option as separate options, this is not efficient in terms of computational complexity. An extension to include quantity as a decision variable should fur-

ther improve the performance. It would be also interesting to study other risk-averse objectives. For example, including the bid values as well as bid quantities as decision variables to the risk-constrained problem formulation in [56] can be considered.

APPENDIX A

APPENDIX OF CHAPTER 2

A.1 Proof of Lemma 1

Let $\tilde{e}_{i,t_i} = \|\hat{x}_{i,t_i} - x_i^*\|_2^2$. By MSPSA update step given in (2.9) and the fact that projection onto Π maps a point closer to x_i^* , we have,

$$\begin{aligned} \tilde{e}_{i,t_i+1} &\leq \left\| \hat{x}_{i,t_i} - a_{t_i} \left(\frac{d_{i,t_i}^+ - d_{i,t_i}^-}{c_{t_i}} \right) \bar{\Delta}_{t_i} - x_i^* \right\|_2^2 \\ &= \tilde{e}_{i,t_i} - 2 \frac{a_{t_i}}{c_{t_i}} (d_{i,t_i}^+ - d_{i,t_i}^-) (\hat{x}_{i,t_i} - x_i^*)^\top \bar{\Delta}_{t_i} + \frac{a_{t_i}^2}{c_{t_i}^2} (d_{i,t_i}^+ - d_{i,t_i}^-)^2 \bar{\Delta}_{t_i}^\top \bar{\Delta}_{t_i}. \end{aligned} \quad (\text{A.1})$$

Our goal is to bound $e_{i,t_i+1} = \mathbb{E}(\tilde{e}_{i,t_i+1} | i, t_i)$ by simplifying (A.1). By (2.6), we obtain,

$$\begin{aligned} \mathbb{E}(d_{i,t_i}^+ - d_{i,t_i}^- | i, t_i, \hat{x}_{i,t_i}, \Delta_{t_i}) &= 4c_{t_i} \Delta_{t_i}^\top \sum_j p_{i,j} A_j^\top (A_j \hat{x}_{i,t_i} + b_j - y^*) \\ &= 4c_{t_i} \Delta_{t_i}^\top \left(\sum_j p_{i,j} A_j^\top A_j \right) (\hat{x}_{i,t_i} - x_i^*) \end{aligned}$$

where last equality is obtained using the FOC for x_i^* . Let $\lambda_{\min,i} = \lambda_{\min}(\sum_j p_{i,j} A_j^\top A_j)$. Using the independence of $\Delta_{t_i,j}$'s, we get,

$$\begin{aligned} -\frac{2}{c_{t_i}} \mathbb{E}((d_{i,t_i}^+ - d_{i,t_i}^-) (\hat{x}_{i,t_i} - x_i^*)^\top \bar{\Delta}_{t_i} | i, t_i, \hat{x}_{i,t_i}) &= -8 \sum_j p_{i,j} \|A_j (\hat{x}_{i,t_i} - x_i^*)\|_2^2 \\ &\leq -8\lambda_{\min,i} \tilde{e}_{i,t_i}. \end{aligned} \quad (\text{A.2})$$

Since Π is compact, $\|\hat{x}_{i,t_i} \pm c_{t_i} \Delta_{t_i}\| \leq \bar{x}$ where constant $\bar{x} = (\max_{x \in \Pi} \|x\|) + \gamma'_i \sqrt{n} \xi_1$. For any $j \in \mathcal{S}$, because b_j and singular values of A_j are bounded, $\|A_j(\hat{x}_{i,t_i} \pm c_{t_i} \Delta_{t_i}) + b_j - y^*\| \leq C_0$ where constant $C_0 = \bar{\sigma} \bar{x} + \bar{b} + \|y^*\|$. By Holder's inequality, we have $\mathbb{E}(\|w_t\|_2^2) \leq m\sigma_w^2$, $\mathbb{E}(\|w_t w_t^\top w_t\|_2) \leq m^2 \sigma_w^3$, and $\mathbb{E}(\|w_t\|_2^4) \leq$

$m^2\sigma_w^4$. Then, after simplification, we obtain, $\mathbb{E}((d_{i,t_i}^\pm)^2|i, t_i, x_{i,t_i}, \Delta_{t_i}) \leq C_1$ where $C_1 = C_0^4 + m^2\sigma_w^4 + 6C_0^2m\sigma_w^2 + 4C_0m^2\sigma_w^3$, and

$$\begin{aligned}\mathbb{E}(-2d_{i,t_i}^+ d_{i,t_i}^-|i, t_i, \hat{x}_{i,t_i}, \Delta_{t_i}) &\leq 8c_{t_i}^2 \left(\Delta_{t_i}^\top \sum_j p_{i,j} A_j^\top (A_j \hat{x}_{i,t_i} + b_j - y^*) \right)^2 \\ &= 8c_{t_i}^2 \left(\Delta_{t_i}^\top \left(\sum_j p_{i,j} A_j^\top A_j \right) (\hat{x}_{i,t_i} - x_i^*) \right)^2,\end{aligned}$$

where last equality is obtained using the FOC of x_i^* . Consequently,

$$\mathbb{E} \left(\left(\frac{d_{i,t_i}^+ - d_{i,t_i}^-}{c_{t_i}} \right)^2 \bar{\Delta}_{t_i}^\top \bar{\Delta}_{t_i} \middle| i, t_i, \hat{x}_{i,t_i} \right) \leq C_2 \tilde{e}_{i,t_i} + \frac{C_3}{c_{t_i}^2}, \quad (\text{A.3})$$

where $C_2 = 8 \max\{2, (1 + (n-1)\xi_1^2\xi_2)\} \bar{\sigma}^4$ and $C_3 = 2C_1 n \xi_2$.

Thus, by expressions (A.1), (A.2), and (A.3);

$$\mathbb{E}(\tilde{e}_{i,t_i+1}|i, t_i, \hat{x}_{i,t_i}) \leq (1 - a_{t_i} 8\lambda_{\min,i} + a_{t_i}^2 C_2) \tilde{e}_{i,t_i} + a_{t_i}^2 \frac{C_3}{c_{t_i}^2}. \quad (\text{A.4})$$

Consequently,

$$e_{i,t_i+1} \leq (1 - a_{t_i} 8\lambda_{\min,i} + a_{t_i}^2 C_2) e_{i,t_i} + a_{t_i}^2 \frac{C_3}{c_{t_i}^2}.$$

Using this result recursively and since $e^x \geq 1 + x$ for all $x \in \mathbb{R}$, we have,

$$\begin{aligned}e_{i,t_i+1} &\leq \left(\prod_{j=1}^{t_i} (1 - 8a_j \lambda_{\min,i} + a_j^2 C_2) \right) e_{i,1} \\ &\quad + \sum_{j=1}^{t_i} \left(\prod_{l=j+1}^{t_i} (1 - 8a_l \lambda_{\min,i} + a_l^2 C_2) \right) a_j^2 \frac{C_3}{c_j^2} \\ &\leq e^{\sum_{j=1}^{t_i} (-8a_j \lambda_{\min,i} + a_j^2 C_2)} e_{i,1} + \sum_{j=1}^{t_i} e^{\sum_{l=j+1}^{t_i} (-8a_l \lambda_{\min,i} + a_l^2 C_2)} a_j^2 \frac{C_3}{c_j^2}.\end{aligned}$$

Since $\gamma_i \geq 1/(8\lambda_{\min,i})$ and $e_{i,1} \leq (2 \max_{x \in \Pi} \|x\|_2)^2$,

$$\begin{aligned}e_{i,t_i+1} &\leq e^{-\log(t_i+1+N_i) + \log(1+N_i) + 2\gamma_i^2 C_2} e_{i,1} + \sum_{j=1}^{t_i} \frac{j + N_i}{t_i + 1 + N_i} e^{(1+N_i)^{-1} + 2\gamma_i^2 C_2} a_j^2 \frac{C_3}{c_j^2} \\ &\leq \frac{C_i'}{t_i + 1 + N_i} + \frac{C_i''}{t_i + 1 + N_i} \sum_{j=1}^{t_i} \frac{1}{\sqrt{j + N_i}} \\ &\leq \frac{C_i}{\sqrt{t_i + 1}},\end{aligned}$$

where $C'_i = (1 + N_i) \exp(2\gamma_i^2 C_2) 4 \max_{x \in \Pi} \|x\|_2^2$, $C''_i = (\gamma_i/\gamma'_i)^2 C_3 \exp((1 + N_i)^{-1} + 2\gamma_i^2 C_2)(1 + (\max\{0, N'_i - N_i\})^{1/2})$, and $C_i = \max\{C'_i, 2C''_i\}$. \square

A.2 Proof of Theorem 2

In the proof of Lemma 1, for any state $i \in \mathcal{S}$, we showed that the inequality (A.4) holds where $\tilde{e}_{i,t_i} = \|\hat{x}_{i,t_i} - x_i^*\|_2^2$. By Theorem 1 of Robbins and Siegmund [69], we know that $\lim_{t_i \rightarrow \infty} \tilde{e}_{i,t_i} < \infty$ exists and $\sum_{t_i=1}^{\infty} 8\lambda_{\min,i} a_{t_i} \tilde{e}_{i,t_i} < \infty$ almost surely (a.s.). Since $\sum_{t_i=1}^{\infty} 8\lambda_{\min,i} a_{t_i} = \infty$, we obtain that

$$\Pr \left(\lim_{t_i \rightarrow \infty} \tilde{e}_{i,t_i} = 0 \right) = 1.$$

Let $1_i(s_t)$ be the indicator function. Given $s_{t-1} = i$ and t_i , x_t^{MSPSA} is equal to either $\hat{x}_{i,t_i} + c_{t_i} \Delta_{t_i}$ or $\hat{x}_{i,t_i} - c_{t_i} \Delta_{t_i}$. Hence, $\|x_t^{\text{MSPSA}} - x_i^*\|_2^2 \leq 2\tilde{e}_{i,t_i} + 2(\gamma'_i)^2 n \xi_1^2 t_i^{-1/2}$. If state i is recurrent, $\Pr(\lim_{t \rightarrow \infty} t_i < \infty) = 0$ because t_i is greater or equal to half of the number of times state i is occurred up to t . Therefore, for a recurrent state $i \in \mathcal{S}$,

$$\begin{aligned} & \Pr \left(\lim_{t \rightarrow \infty} 1_i(s_{t-1}) \|x_t^{\text{MSPSA}} - x_i^*\|_2^2 = 0 \right) \\ &= \Pr \left(\lim_{t \rightarrow \infty} 1_i(s_{t-1}) \|x_t^{\text{MSPSA}} - x_i^*\|_2^2 = 0 \mid \lim_{t \rightarrow \infty} t_i = \infty \right) \\ &\geq \Pr \left(\lim_{t \rightarrow \infty} (2\tilde{e}_{i,t_i} + 2(\gamma'_i)^2 n \xi_1^2 t_i^{-1/2}) = 0 \mid \lim_{t \rightarrow \infty} t_i = \infty \right) \\ &= \Pr \left(\lim_{t_i \rightarrow \infty} \tilde{e}_{i,t_i} = 0 \right) \\ &= 1. \end{aligned}$$

So, for a recurrent state i and for any $\epsilon > 0$, we have,

$$\lim_{t' \rightarrow \infty} \Pr \left(1_i(s_{t-1}) \|x_t^{\text{MSPSA}} - x_i^*\|_2^2 > \epsilon \text{ for some } t \geq t' \right) = 0. \quad (\text{A.5})$$

If a state $i \in \mathcal{S}$ is transient, then for any $\epsilon > 0$, we have,

$$\begin{aligned}
& \lim_{t' \rightarrow \infty} \Pr \left(1_i(s_{t-1}) \|x_t^{\text{MSPSA}} - x_i^*\|_2^2 > \epsilon \text{ for some } t \geq t' \right) \\
& \leq \lim_{t' \rightarrow \infty} \Pr (s_{t-1} = i \text{ for some } t \geq t') \\
& = 0,
\end{aligned} \tag{A.6}$$

where last equality is due to Borel-Cantelli lemma and the fact that $\sum_{t=0}^{\infty} \Pr(s_t = i) < \infty$ for a transient state i .

By definition, expression (2.11) holds, if and only if, for every $\epsilon > 0$, $\lim_{t' \rightarrow \infty} \Pr(\|x_t^{\text{MSPSA}} - x_{s_{t-1}}^*\|_2^2 > \epsilon \text{ for some } t \geq t') = 0$. Any state $i \in \mathcal{S}$ is either recurrent or transient. Hence, by (A.5) and (A.6), we obtain that, for any $\epsilon > 0$,

$$\begin{aligned}
& \lim_{t \rightarrow \infty} \Pr \left(\|x_t^{\text{MSPSA}} - x_{s_{t-1}}^*\|_2^2 > \epsilon \text{ for some } t \geq t' \right) \\
& \leq \lim_{t \rightarrow \infty} \sum_{i=1}^K \Pr \left(1_i(s_{t-1}) \|x_t^{\text{MSPSA}} - x_i^*\|_2^2 > \epsilon \text{ for some } t \geq t' \right) \\
& = 0.
\end{aligned}$$

Since (2.11) holds and $\|x_T^{\text{MSPSA}} - x_i^*\|_2^2 \leq C_0$ where $C_0 = (2 \max_{x \in \Pi} \|x\| + \gamma'_i \sqrt{n} \xi_1)^2$, by Lebesgue's dominated convergence theorem,

$$\lim_{t \rightarrow \infty} \mathbb{E} \left(\|x_t^{\text{MSPSA}} - x_{s_{t-1}}^*\|_2^2 \right) = 0.$$

□

A.3 Proof of Theorem 3

Let the transition probability from any state to any other state be $1/K$. Without loss of generality, take $y^* = 0$. Let w_t be i.i.d. with distribution $N(\mathbf{0}_m, \sigma_w^2 I_m)$

which is independent of the state.

Because additional observations can't increase the growth rate of regret for an optimal policy, we assume that the decision maker receives the observation values corresponding to the input x_t^μ from all other states that didn't occur at time t as additional observations at time t . Hence, at each t , the decision maker gets observations from the affine functions of all states for input x_t^μ . Let's define A , b , and w_t as

$$A = \begin{bmatrix} A_1 \\ \vdots \\ A_K \end{bmatrix} \quad b = \begin{bmatrix} b_1 \\ \vdots \\ b_K \end{bmatrix} \quad w_t = \begin{bmatrix} w_t^{(1)} \\ \vdots \\ w_t^{(K)} \end{bmatrix}$$

where $w_t^{(i)}$ denotes the system noise of observation from state i . Now, for any policy μ , we can express the observation vector at t as

$$y_t^\mu = Ax_t^\mu + b + w_t.$$

Observe that FOC for the optimal input $x_{s_{t-1}}^*$ at time t obtained from minimizing (2.6) is the same for any state $s_{t-1} \in \mathcal{S}$ for our fixed choice of P . Hence, we drop the dependence on the previous state s_{t-1} along with P and denote it as $x^*(\theta)$, i.e., $x_{s_{t-1}}^* = x^*(\theta)$, to express the dependence on θ . With the new notation, FOC can be expressed as

$$\nu = A^\top(Ax^*(\theta) + b) = 0.$$

Consequently, the optimal price given in (2.7) becomes

$$x^*(\theta) = -(A^\top A)^{-1} A^\top b.$$

Let's express θ_k as $\theta_k = [b_{k,1}, a_{k,1}, \dots, b_{k,m}, a_{k,m}]^\top$ where $b_{k,i}$ is the i th entry of b_k and $a_{k,i}$ is the i th row vector of A_k . We fix a compact rectangle $\Theta \subset \mathbb{R}^{K \times m \times (n+1)}$

such that, for any $\theta \in \Theta$, $x^*(\theta)$ is contained in Π and A_k is full column rank for all $k \in \mathcal{S}$. * Since P and $\{f_i\}_{i=1}^K$ are already fixed, our goal is to obtain the performance of the worst-case system parameter θ that is chosen from the set Θ .

Applying implicit function theorem on ν gives

$$\frac{\partial x^*(\theta)}{\partial \theta} = - \left(\frac{\partial \nu}{\partial x^*(\theta)} \right)^{-1} \frac{\partial \nu}{\partial \theta} = - (A^\top A)^{-1} \frac{\partial \nu}{\partial \theta}, \quad (\text{A.7})$$

and by calculus, we have,

$$\left(\frac{\partial \nu}{\partial \theta} \right)^\top = (Ax^*(\theta) + b) \otimes \begin{bmatrix} \mathbf{0}_n^\top \\ \mathbb{I}_n \end{bmatrix} + A \otimes \begin{bmatrix} 1 \\ x^*(\theta) \end{bmatrix}. \quad (\text{A.8})$$

Let $M = Km$. Density of the output vector up to time t given the parameter vector θ and input vector X^t can be written as

$$g(Y^t|X^t, \theta) = \prod_{i=1}^t \frac{\exp(-\|y_i^\mu - b - Ax_i^\mu\|_2^2 / (2\sigma_w^2))}{(2\pi\sigma_w^2)^{M/2}}.$$

By writing the joint distribution as a product of conditionals and by the conditional independence of the input for any policy μ from the parameter θ given the information history vector I_{t-1} , we get,

$$\frac{\partial \log g(Y^t, X^t|\theta)}{\partial \theta} = \frac{\partial \log g(Y^t|X^t, \theta)}{\partial \theta} = \frac{1}{\sigma_w^2} \sum_{i=1}^t w_i \otimes \begin{bmatrix} 1 \\ x_i^\mu \end{bmatrix}.$$

By using the mixed product property $(A \otimes B)(C \otimes D) = AC \otimes BD$ and the

The existence of Θ can be shown by the continuity of $x^(\theta)$ on a compact rectangle Θ' which satisfies A_k to be full column rank for all $k \in \mathcal{S}$ and for any $\theta \in \Theta'$, and contains a fixed point θ' in its interior for which $x^*(\theta')$ is in the interior of Π . The existence of Θ' can be shown by using the continuity of the determinant of $A_k^\top A_k$ for each $k \in \mathcal{S}$ at the fixed point θ' .

independence of w_i , we obtain the fisher information for g as

$$\begin{aligned} I_t^\mu(\theta) &= \mathbb{E} \left(\frac{\partial \log g(Y^t|X^t, \theta)}{\partial \theta} \frac{\partial \log g(Y^t|X^t, \theta)^\top}{\partial \theta} \middle| X^t, \theta \right) \\ &= \frac{1}{\sigma_w^2} I_M \otimes \left(\sum_{i=1}^t \begin{bmatrix} 1 \\ x_i^\mu \end{bmatrix} \begin{bmatrix} 1, (x_i^\mu)^\top \end{bmatrix} \right). \end{aligned} \quad (\text{A.9})$$

Now, we choose a prior distribution λ as an absolutely continuous density on Θ taking positive values in the interior of Θ and zero on its boundary. We choose A and b to be independently distributed with distributions λ_A and λ_b , respectively, so that $\lambda = \lambda_A \lambda_b$. Take $C(\theta) = b^\top \otimes \begin{bmatrix} -x^*(\theta), I_n \end{bmatrix}$. Now, we use the multivariate van Trees inequality [35] in a similar way in [42]. This inequality can be expressed as

$$\mathbb{E} (\|\hat{x}_t^\mu - x^*(\theta)\|_2^2) \geq \frac{\left(\mathbb{E} \left(\text{tr} \left(C(\theta) \frac{\partial x^*(\theta)^\top}{\partial \theta} \right) \right) \right)^2}{\mathbb{E} \left(\text{tr} \left(C(\theta) I_{t-1}^\mu(\theta) C(\theta)^\top \right) \right) + \tilde{I}(\lambda)} \quad (\text{A.10})$$

where the expectation operators are also taken over the prior distribution λ and $\tilde{I}(\lambda)$ is some constant given λ , which can be seen as the Fisher information for the distribution λ .

By (A.9), we have,

$$\begin{aligned} \text{tr} \left(C(\theta) I_{t-1}^\mu(\theta) C(\theta)^\top \right) &= \frac{b^\top b}{\sigma_w^2} \sum_{i=1}^{t-1} \|x_i^\mu - x^*(\theta)\|_2^2 \\ &\leq c_0 \sum_{i=1}^{t-1} \|x_i^\mu - x^*(\theta)\|_2^2 \end{aligned} \quad (\text{A.11})$$

where $c_0 = (K\bar{b}^2)/\sigma_w^2$.

Let's define $P = I_M - A(A^\top A)^{-1}A^\top$. Since $A^\top A$ is symmetric positive definite,

by (A.7) and (A.8), we obtain

$$\begin{aligned}
\text{tr} \left(C(\theta) \frac{\partial x^*(\theta)^\top}{\partial \theta} \right) &= \text{tr} \left(-C(\theta) \frac{\partial \nu^\top}{\partial \theta} (A^\top A)^{-1} \right) \\
&= -b^\top (Ax^*(\theta) + b) \text{tr} \left((A^\top A)^{-1} \right) \\
&= -b^\top P b \text{tr} \left((A^\top A)^{-1} \right).
\end{aligned}$$

By singular value decomposition (SVD) of A , observe that $P = UD(\mathbf{0}_n^\top, \mathbf{1}_{M-n}^\top)U^\top$ where $U \in \mathbb{R}^{M \times M}$ is an orthogonal matrix, and $D(d_1, \dots, d_M)$ denotes a diagonal matrix with diagonal entries d_1, \dots, d_M . Hence, P is symmetric positive semidefinite. Also observe that $\text{tr}((A^\top A)^{-1}) \geq n/(K\bar{\sigma}^2)$. Then, we can bound the numerator term,

$$\left(\mathbb{E} \left(\text{tr} \left(C(\theta) \frac{\partial x^*(\theta)^\top}{\partial \theta} \right) \right) \right)^2 \geq \frac{n^2}{K^2 \bar{\sigma}^4} (\mathbb{E}(b^\top P b))^2. \quad (\text{A.12})$$

Observe that $\bar{P} = \mathbb{E}(P)$ is symmetric positive semidefinite and nonzero for $K > 1$ (or $K = 1$ and $m > n$) since $\text{tr}(\bar{P}) = \mathbb{E}(\text{tr}(P)) = Km - n$. Hence, there exists some direction $z \in \mathbb{R}^M$ such that $z^\top \bar{P} z > 0$, and, consequently, there exists some distribution λ_b such that $\mathbb{E}(b)^\top \bar{P} \mathbb{E}(b) > 0$. More specifically, if $\mathbb{E}(b)^\top \bar{P} \mathbb{E}(b) = 0$ for some choice of λ_b , we can change that choice of λ_b to shift the mean of b slightly in the direction of z , and have $\mathbb{E}(b)^\top \bar{P} \mathbb{E}(b) > 0$. By independence of b and P ,

$$\begin{aligned}
\mathbb{E}(b^\top P b) &= \mathbb{E}(b)^\top \bar{P} \mathbb{E}(b) + \mathbb{E}((b - \mathbb{E}(b))^\top P (b - \mathbb{E}(b))) \\
&\geq \mathbb{E}(b)^\top \bar{P} \mathbb{E}(b).
\end{aligned} \quad (\text{A.13})$$

Hence, by expressions (A.10), (A.11), (A.12), and (A.13);

$$\begin{aligned} \sum_{t=2}^T \mathbb{E} (\|\hat{x}_t^\mu - x^*(\theta)\|_2^2) &\geq \sum_{t=2}^T \frac{c_1}{\mathbb{E}(\sum_{i=1}^{t-1} \|x_i^\mu - x^*(\theta)\|_2^2) + c_2} \\ &\geq \sum_{t=2}^T \frac{c_1}{\mathbb{E}(\sum_{i=1}^T \|x_i^\mu - x^*(\theta)\|_2^2) + c_2}, \end{aligned} \tag{A.14}$$

where $c_1 = n^2(\mathbb{E}(b)^\top \bar{P} \mathbb{E}(b))^2 / (K^2 \bar{\sigma}^4 c_0)$ and $c_2 = \tilde{I}(\lambda) / c_0$.

Since $\bar{\mathcal{E}}_T^\mu \geq \mathbb{E}(\sum_{i=1}^T \|x_i^\mu - x^*(\theta)\|_2^2)$, by (A.14), we have,

$$\hat{\mathcal{E}}_T^\mu \geq \frac{c_1(T-1)}{\bar{\mathcal{E}}_T^\mu + c_2} \geq \frac{c_1(T-1)}{(1 + c_2/\bar{\mathcal{E}}_1^\mu) \bar{\mathcal{E}}_T^\mu}.$$

Let $x_k^*(\theta)$ denote the k th entry of $x^*(\theta)$, and, by extreme value theorem, $u_k = \sup_{\theta \in \Theta} x_k^*(\theta)$ and $l_k = \inf_{\theta \in \Theta} x_k^*(\theta)$ are attained. Since $x^*(\theta)$ is not a constant over Θ (otherwise $\partial x^*(\theta) / \partial \theta$ would be zero for all $\theta \in \Theta$, and left hand side of (A.12) would be zero for any λ which is a contradiction), $\max_{k \in \{1, \dots, n\}} (u_k - l_k) > 0$. For any policy μ , $\bar{\mathcal{E}}_1^\mu = \sup_{\theta \in \Theta} \mathbb{E}(\|x_1^\mu - x^*(\theta)\|_2^2 | \theta) \geq \max_{k \in \{1, \dots, n\}} ((u_k - l_k)/2)^2 > 0$. Hence, we have, $\hat{\mathcal{E}}_T^\mu \geq (CT) / \bar{\mathcal{E}}_T^\mu$ where $C = c_1/2 / (1 + (4c_2 / \max_{k \in \{1, \dots, n\}} (u_k - l_k)^2))$. \square

A.4 Proof of Lemma 2

We will follow the steps in Lemma 1 and simplify inequality (A.1). By (2.16), we obtain,

$$\begin{aligned} \mathbb{E} (d_{i,t_i}^+ - d_{i,t_i}^- | i, t_i, \hat{x}_{i,t_i}, \Delta_{t_i}) &= -2c_{t_i} \Delta_{t_i}^\top \sum_j p_{i,j} ((A_j + A_j^\top) \hat{x}_{i,t_i} + b_j) \\ &= -2c_{t_i} \Delta_{t_i}^\top \left(\sum_j p_{i,j} (A_j + A_j^\top) \right) (\hat{x}_{i,t_i} - x_i^*), \end{aligned}$$

where last equality is obtained using the FOC for x_i^* . Let $\lambda_{min,i} = \lambda_{min}(-\sum_j p_{i,j}(A_j + A_j^\top)/2)$. Using the independence of $\Delta_{t_i,j}$'s, we obtain,

$$\begin{aligned} & -\frac{2}{c_{t_i}} \mathbb{E} \left((d_{i,t_i}^+ - d_{i,t_i}^-) (\hat{x}_{i,t_i} - x_i^*)^\top \bar{\Delta}_{t_i} \middle| i, t_i, \hat{x}_{i,t_i} \right) \\ & = 4(\hat{x}_{i,t_i} - x_i^*)^\top \left(\sum_j p_{i,j}(A_j + A_j^\top) \right) (\hat{x}_{i,t_i} - x_i^*) \\ & \leq -8\lambda_{min,i} \tilde{e}_{i,t_i}. \end{aligned}$$

As in Lemma 1, $\|\hat{x}_{i,t_i} \pm c_{t_i} \Delta_{t_i}\| \leq \bar{x}$. Hence, for any $j \in \mathcal{S}$, $(\hat{x}_{i,t_i} \pm c_{t_i} \Delta_{t_i})^\top (A_j(\hat{x}_{i,t_i} \pm c_{t_i} \Delta_{t_i}) + b_j) \leq \bar{\sigma} \bar{x}^2 + \bar{b} \bar{x}$. Since $\mathbb{E}(\|w_t\|_2^2) \leq n\sigma_w^2$, $\mathbb{E}((d_{i,t_i}^\pm)^2 | i, t_i, \hat{x}_{i,t_i}, \Delta_{t_i}) \leq C_1$ where constant $C_1 = (\bar{\sigma} \bar{x}^2 + \bar{b} \bar{x})^2 + n\sigma_w^2 \bar{x}^2$. Consequently,

$$\mathbb{E} \left((d_{i,t_i}^+ - d_{i,t_i}^-)^2 \middle| i, t_i, \hat{x}_{i,t_i}, \Delta_{t_i} \right) \leq 2c_{t_i}^2 \left(\Delta_{t_i}^\top \left(\sum_j p_{i,j}(A_j + A_j^\top) \right) (\hat{x}_{i,t_i} - x_i^*) \right)^2 + 2C_1,$$

and thus, we obtain (A.3) where $C_2 = 8 \max\{2, (1 + (n-1)\xi_1^2 \xi_2)\} \bar{\sigma}^2$ and $C_3 = 2C_1 n \xi_2$. Therefore, (A.4) holds and the rest of the proof is the same as in Lemma 1. \square

A.5 Proof of Theorem 7

The inequality given in (2.19) is used to obtain (2.20) and (2.21) as in Theorem 4. The proof of inequality (2.19) follows the proof of Theorem 3 with some slight modifications to bound the numerator term of the van Trees inequality due to revenue maximization objective.

The FOC for $x^*(\theta)$ becomes $\nu = \sum_j ((A_j^\top + A_j)x^*(\theta) + b_j) = 0$ and the optimal price $x^*(\theta) = (\sum_j -(A_j^\top + A_j))^{-1}(\sum_j b_j)$. For this problem, the compact rectangle

Θ is such that, for any $\theta \in \Theta$, $x^*(\theta)$ is contained in Π and A_k is negative definite for all $k \in \mathcal{S}$.^{*} The implicit function theorem on ν gives

$$\frac{\partial x^*(\theta)}{\partial \theta} = - \left(\sum_j (A_j^\top + A_j) \right)^{-1} \frac{\partial \nu}{\partial \theta}$$

where

$$\left(\frac{\partial \nu}{\partial \theta} \right)^\top = \mathbf{1}_K \otimes \left(x^*(\theta) \otimes \begin{bmatrix} \mathbf{0}_n^\top \\ \mathbb{I}_n \end{bmatrix} + \mathbb{I}_n \otimes \begin{bmatrix} 1 \\ x^*(\theta) \end{bmatrix} \right).$$

We take λ_b such that $\mathbb{E}(\sum_j b_j) \neq 0$. Consequently, the numerator term of the van Trees Inequality can be bounded as

$$\begin{aligned} \mathbb{E} \left(\text{tr} \left(\mathbb{C}(\theta) \frac{\partial x^*(\theta)}{\partial \theta} \right)^\top \right) &= \mathbb{E} \left(\left(\sum_j b_j \right)^\top x^*(\theta) \text{tr} \left(\sum_j - (A_j^\top + A_j) \right)^{-1} \right) \\ &\geq \frac{n \mathbb{E} \left(\left(\sum_j b_j \right)^\top \left(\sum_j b_j \right) \right)}{(2K\bar{\sigma})^2} \\ &\geq \frac{n \mathbb{E} \left(\sum_j b_j \right)^\top \mathbb{E} \left(\sum_j b_j \right)}{(2K\bar{\sigma})^2} \\ &> 0. \end{aligned}$$

□

^{*}The existence of such a set can shown by the same argument as in Theorem 3 by using the continuity of the maximum eigenvalue of A_k rather than the determinant.

APPENDIX B

APPENDIX OF CHAPTER 3

B.1 Proof of Theorem 8

Recall that $x^* = \arg \max_{x \in \mathcal{F}} r^{(\rho)}(x)$ and let $x_{t+1}^* = \arg \max_{x \in \mathcal{F}_t} r^{(\rho)}(x)$. Hence, for any $x' \in \mathcal{F}_t$,

$$r^{(\rho)}(x^*) - r^{(\rho)}(x_{t+1}^*) \leq r^{(\rho)}(x^*) - r^{(\rho)}(x').$$

Let x_k^* and x'_k denote the k th entry of x^* and x' , respectively. We take $x'_k = \lfloor x_k^*/(B/\alpha_t) \rfloor (B/\alpha_t)$ for all $k \in \{1, \dots, K\}$, where $\lfloor x_k^*/(B/\alpha_t) \rfloor$ denotes the largest integer smaller or equal to $x_k^*/(B/\alpha_t)$, so that $x' \in \mathcal{F}_t$ and $|x'_k - x_k^*| \leq B/\alpha_t$ for all $k \in \{1, \dots, K\}$. Then, due to Lipschitz continuity of $r^{(\rho)}(\cdot)$ on \mathcal{F} with p-norm and constant L ,

$$r^{(\rho)}(x^*) - r^{(\rho)}(x_{t+1}^*) \leq LK^{1/p}B/\alpha_t. \quad (\text{B.1})$$

Since the payoff obtained at each period t from bidding on a node $k \in \{1, \dots, K\}$ is in $[l, u]$ and $r^{(\rho)}(\cdot)$ is Lipschitz, $r^{(\rho)}(x_{t+1}^*) - r^{(\rho)}(x) \leq c_1$ for any $x \in \mathcal{F}_t$ where $c_1 = \min(c_2, LK^{1/p}B)$ and $c_2 = K((u-l) + \rho(u-l)^2)$. Then, for any $\delta_t > 0$,

$$\begin{aligned} r^{(\rho)}(x_{t+1}^*) - r^{(\rho)}(x_{t+1}^{\text{DPDS}}) &= \sum_{x \in \mathcal{F}_t} (r^{(\rho)}(x_{t+1}^*) - r^{(\rho)}(x)) \mathbb{1}\{x_{t+1}^{\text{DPDS}} = x\} \\ &\leq \delta_t \sum_{x \in \mathcal{F}_t: r^{(\rho)}(x_{t+1}^*) - r^{(\rho)}(x) \leq \delta_t} \mathbb{1}\{x_{t+1}^{\text{DPDS}} = x\} \\ &\quad + c_1 \sum_{x \in \mathcal{F}_t: r^{(\rho)}(x_{t+1}^*) - r^{(\rho)}(x) > \delta_t} \mathbb{1}\{x_{t+1}^{\text{DPDS}} = x\} \\ &\leq \delta_t + c_1 \sum_{x \in \mathcal{F}_t: r^{(\rho)}(x_{t+1}^*) - r^{(\rho)}(x) > \delta_t} \mathbb{1}\{x_{t+1}^{\text{DPDS}} = x\} \end{aligned}$$

where the last inequality is obtained by the fact that at most one of the indicator functions can be equal to one due to the events being disjoint.

Since DPDS chooses $x_{t+1} \in \mathcal{F}_t$ that maximizes $\bar{r}_t^{(\rho)}(x_{t+1}) = \sum_{k=1}^K \bar{r}_{t,k}^{(\rho)}(x_{t+1,k})$, $\bar{r}_t^{(\rho)}(x) \geq \bar{r}_t^{(\rho)}(x_{t+1}^*)$ has to hold for any $x \in \mathcal{F}_t$ if $x_{t+1}^{\text{DPDS}} = x$. Hence, we can upper bound the last inequality obtained to get

$$r^{(\rho)}(x_{t+1}^*) - r^{(\rho)}(x_{t+1}^{\text{DPDS}}) \leq \delta_t + c_1 \sum_{x \in \mathcal{F}_t: r^{(\rho)}(x_{t+1}^*) - r^{(\rho)}(x) > \delta_t} \mathbb{1}\{\bar{r}_t^{(\rho)}(x) \geq \bar{r}_t^{(\rho)}(x_{t+1}^*)\}.$$

In order for $\bar{r}_t^{(\rho)}(x) \geq \bar{r}_t^{(\rho)}(x_{t+1}^*)$ to hold for any $x \in \mathcal{F}_t$ satisfying $r^{(\rho)}(x_{t+1}^*) - r^{(\rho)}(x) > \delta_t$, observe that the event

$$\mathcal{E}_1 = \left\{ \bar{r}_t^{(\rho)}(x_{t+1}^*) + \delta_t/2 \leq r^{(\rho)}(x_{t+1}^*) \right\}$$

holds and/or the event

$$\mathcal{E}_2 = \left\{ r^{(\rho)}(x) + \delta_t/2 \leq \bar{r}_t^{(\rho)}(x) \right\}$$

holds. Consequently,

$$\mathbb{E} \left(r^{(\rho)}(x_{t+1}^*) - r^{(\rho)}(x_{t+1}^{\text{DPDS}}) \right) \leq \delta_t + c_1 \sum_{x \in \mathcal{F}_t: r^{(\rho)}(x_{t+1}^*) - r^{(\rho)}(x) > \delta_t} \Pr(\mathcal{E}_1 \cup \mathcal{E}_2).$$

Also, observe that, for any fixed $x \in \mathcal{F}$, $\mathbb{E} \left(\bar{r}_t^{(\rho)}(x) \middle| x \right) = r^{(\rho)}(x)$. So, we can use McDiarmid's inequality [62] to upper bound both $\Pr(\mathcal{E}_1)$ and $\Pr(\mathcal{E}_2)$ if we show that $\bar{r}_t^{(\rho)}(x)$ for fixed $x \in \mathcal{F}_t$ satisfies the bounded differences condition as a function of $\{(\lambda_i, \pi_i)\}_{i=1}^t \in \Pi^t$ where Π denotes the support of the random variable (λ_i, π_i) .

Let x_k denote the k th entry of x . Define $\bar{r}_t^{(-j)}(x) = \sum_{k=1}^K \bar{r}_{t,k}^{(-j)}(x_k)$ where

$$\bar{r}_{t,k}^{(-j)}(x_k) = \frac{1}{t-1} \sum_{i: i \neq j, 1 \leq i \leq t} (\pi_{i,k} - \lambda_{i,k}) \mathbb{1}\{x_k \geq \lambda_{i,k}\},$$

and define $\bar{v}_t^{(-j)}(x) = \sum_{k=1}^K \bar{v}_{t,k}^{(-j)}(x_k)$ where

$$\bar{v}_{t,k}^{(-j)}(x_k) = \frac{1}{t-1} \sum_{i:i \neq j, 1 \leq i \leq t} \left((\pi_{i,k} - \lambda_{i,k}) \mathbb{1}\{x_k \geq \lambda_{i,k}\} - \bar{r}_{t,k}^{(-j)}(x_k) \right)^2.$$

Then, for any $j \in \{1, \dots, t\}$, we can express $\bar{r}_t^{(\rho)}(x)$ as follows:

$$\begin{aligned} \bar{r}_t^{(\rho)}(x) &= h_t^{(-j)}(x) + \frac{1}{t} \sum_{k=1}^K (\pi_{j,k} - \lambda_{j,k}) \mathbb{1}\{x_k \geq \lambda_{j,k}\} \\ &\quad - \frac{\rho}{t} \sum_{k=1}^K \left((\pi_{j,k} - \lambda_{j,k}) \mathbb{1}\{x_k \geq \lambda_{j,k}\} - \bar{r}_{t,k}^{(-j)}(x_k) \right)^2 \end{aligned}$$

where

$$h_t^{(-j)}(x) = \frac{t-1}{t} \bar{r}_t^{(-j)}(x) - \rho \bar{v}_t^{(-j)}(x)$$

doesn't depend on (λ_j, π_j) . We also define $\bar{r}_t^{(\rho, j')}(x)$ as the average payoff function that would result from observing $(\lambda_{j'}, \pi_{j'})$ instead of (λ_j, π_j) at period j .

Consequently,

$$\begin{aligned} \bar{r}_t^{(\rho, j')}(x) &= h_t^{(-j)}(x) + \frac{1}{t} \sum_{k=1}^K (\pi_{j',k} - \lambda_{j',k}) \mathbb{1}\{x_k \geq \lambda_{j',k}\} \\ &\quad - \frac{\rho}{t} \sum_{k=1}^K \left((\pi_{j',k} - \lambda_{j',k}) \mathbb{1}\{x_k \geq \lambda_{j',k}\} - \bar{r}_{t,k}^{(-j)}(x_k) \right)^2. \end{aligned}$$

Recall that, for any $(\lambda_i, \pi_i) \in \Pi$, $x \in \mathcal{F}$ and $k \in \{1, \dots, K\}$, $l \leq (\pi_{i,k} - \lambda_{i,k}) \mathbb{1}\{x_k \geq \lambda_{i,k}\} \leq u$. Therefore, for any $j \in \{1, \dots, t\}$ and $x \in \mathcal{F}$, $\bar{r}_t^{(\rho)}(x), \bar{r}_t^{(\rho, j')}(x) \in \left[h_t^{(-j)}(x) + K(l - \rho(u - l)^2)/t, h_t^{(-j)}(x) + Ku/t \right]$ for any $\{(\lambda_i, \pi_i)\}_{i=1}^t, (\lambda_{j'}, \pi_{j'}) \in \Pi^{t+1}$. Hence, for any $x \in \mathcal{F}$ and $j \in \{1, \dots, t\}$,

$$\sup_{\{(\lambda_i, \pi_i)\}_{i=1}^t, (\lambda_{j'}, \pi_{j'}) \in \Pi^{t+1}} \left| \bar{r}_t^{(\rho)}(x) - \bar{r}_t^{(\rho, j')}(x) \right| \leq \frac{c_2}{t}.$$

Since bounded differences condition holds, by McDiarmid's inequality, both $\Pr(\mathcal{E}_1)$ and $\Pr(\mathcal{E}_2)$ are upper bounded by $\exp(-t\delta_t^2/(2c_2^2))$. Using the fact that

the cardinality of the set $\{x \in \mathcal{F}_t : r^{(\rho)}(x_{t+1}^*) - r^{(\rho)}(x) > \delta_t\}$ is upper bounded by $\alpha_t^K + K \leq 2\alpha_t^K$ for $\alpha_t \geq 2$ and $\Pr(\mathcal{E}_1 \cup \mathcal{E}_2) \leq \Pr(\mathcal{E}_1) + \Pr(\mathcal{E}_2)$, we get

$$\mathbb{E} \left(r^{(\rho)}(x_{t+1}^*) - r^{(\rho)}(x_{t+1}^{\text{DPDS}}) \right) \leq \delta_t + 4c_1\alpha_t^K \exp \left(-\frac{t\delta_t^2}{2c_2^2} \right). \quad (\text{B.2})$$

By setting $\delta_t = c_2\sqrt{2(\gamma+1)K+1}\sqrt{\log t/t}$ and $\alpha_t = \max(\lceil \alpha t^\gamma \rceil, 2)$ with $\gamma \geq 1/2$ and $\alpha > 0$, from (B.1) and (B.2), we obtain

$$\begin{aligned} \mathbb{E}(r^{(\rho)}(x^*) - r^{(\rho)}(x_{t+1}^{\text{DPDS}})) &= \mathbb{E}(r^{(\rho)}(x^*) - r^{(\rho)}(x_{t+1}^*)) + \mathbb{E}(r^{(\rho)}(x_{t+1}^*) - r^{(\rho)}(x_{t+1}^{\text{DPDS}})) \\ &\leq LK^{1/p}B/\alpha_t + C_1\sqrt{\log t/t} + 4c_1\alpha_t^K t^{-(\gamma+1)K-1/2} \\ &\leq (LK^{1/p}B/\alpha + 4c_1 \max(t^{-K/2}, ((\alpha+1)/t)^K)) t^{-1/2} \\ &\quad + C_1\sqrt{\log t/t} \\ &\leq C_1\sqrt{\log t/t} + C_2t^{-1/2}, \end{aligned}$$

where $C_1 = c_2\sqrt{2(\gamma+1)K+1}$ and $C_2 = LK^{1/p}B/\alpha + 4c_1 \max(1, \alpha^K)$.

For any $T \geq 2$, $\sum_{t=1}^{T-1} 1/\sqrt{t} \leq 2\sqrt{T-1} - 1$ and $\sum_{t=1}^{T-1} \sqrt{\log t/t} \leq 2\sqrt{(T-1)\log(T-1)}$. Hence, for $T > 2$,

$$\begin{aligned} \sum_{t=2}^{T-1} \mathbb{E} \left(r^{(\rho)}(x^*) - r^{(\rho)}(x_{t+1}^{\text{DPDS}}) \right) &\leq C_1 \sum_{t=1}^{T-1} \sqrt{\frac{\log t}{t}} + C_2 \sum_{t=1}^{T-1} \frac{1}{\sqrt{t}} \\ &\leq 2C_1\sqrt{(T-1)\log(T-1)} + C_2 \left(2\sqrt{T-1} - 1 \right). \end{aligned}$$

Since $\mathbb{E} \left(r^{(\rho)}(x^*) - r^{(\rho)}(x_t^{\text{DPDS}}) \right) \leq c_1$, for any $T \geq 1$,

$$\mathcal{R}_T^{\text{DPDS}}(f) \leq 2C_1\sqrt{T\log T} + 2C_2\sqrt{T}$$

and for any $T > 1$,

$$\mathcal{R}_T^{\text{DPDS}}(f) \leq 2(C_1 + C_2)\sqrt{T\log T}.$$

□

B.2 Proof of Theorem 9

Let λ_t and π_t be independent random variables with distributions

$$f_\lambda(\lambda_t) = \epsilon^{-1} \mathbb{1}\{(1 - \epsilon)/2 \leq \lambda_t \leq (1 + \epsilon)/2\}$$

and $f_\pi(\pi_t) = \text{Bernoulli}(\bar{\pi})$, respectively. Let $f(\lambda_t, \pi_t) = f_\lambda(\lambda_t)f_\pi(\pi_t)$ and $\epsilon = T^{-1/2}/2\sqrt{5}$.

Fix any policy μ . Since λ_t and π_t are independent,

$$r^{(0)}(x) = \mathbb{E}((\bar{\pi} - \lambda_t) \mathbb{1}\{x \geq \lambda_t\} | x)$$

and

$$r^{(0)}(x^*) - r^{(0)}(x_t^\mu) = \mathbb{E}((\bar{\pi} - \lambda_t)(\mathbb{1}\{x^* \geq \lambda_t\} - \mathbb{1}\{x_t^\mu \geq \lambda_t\}) | x_t^\mu, x^*) \quad (\text{B.3})$$

Let f_0, f_1, f_2 denote the distribution of $\{\lambda_t, \pi_t\}_{t=1}^T$ and policy μ under the choice of $\bar{\pi} = 1/2$, $\bar{\pi} = 1/2 - \epsilon$, and $\bar{\pi} = 1/2 + \epsilon$, respectively. Also, let $\mathbb{E}_i(\cdot)$ and $\mathcal{R}_T^\mu(f_i)$ denote the expectation with respect to the distribution f_i and the regret of policy μ under distribution f_i , respectively.

Under distribution f_1 , observe that $\bar{\pi} - \lambda_t \leq -\epsilon/2$ for any value of λ_t . Therefore, optimal solution under known distribution $x^* \in [0, (1 - \epsilon)/2]$ so that $\mathbb{1}\{x^* \geq \lambda_t\} = 0$. Then, by (B.3), the regret can be expressed as

$$\begin{aligned} \mathcal{R}_T^\mu(f_1) &= \mathbb{E}_1 \left(\sum_{t=1}^T -(\bar{\pi} - \lambda_t) \mathbb{1}\{x_t^\mu \geq \lambda_t\} \right) \\ &\geq \frac{\epsilon}{2} \mathbb{E}_1 \left(\sum_{t=1}^T \mathbb{1}\{x_t^\mu \geq \lambda_t\} \right). \end{aligned}$$

Similarly, under distribution f_2 , observe that $\bar{\pi} - \lambda_t \geq \epsilon/2$ for any value of λ_t . Therefore, optimal solution under known distribution $x^* \in [(1 + \epsilon)/2, 1]$ so that

$\mathbb{1}\{x^* \geq \lambda_t\} = 1$. Then, by (B.3), the regret becomes

$$\begin{aligned}\mathcal{R}_T^\mu(f_2) &= \mathbb{E}_2 \left(\sum_{t=1}^T (\bar{\pi} - \lambda_t) \mathbb{1}\{x_t^\mu < \lambda_t\} \right) \\ &\geq \frac{\epsilon}{2} \mathbb{E}_2 \left(\sum_{t=1}^T \mathbb{1}\{x_t^\mu < \lambda_t\} \right).\end{aligned}$$

For any non-negative bounded function h defined on information history $I_T = \{x_t, \lambda_t, \pi_t\}_{t=1}^T$ such that $0 \leq h(I_T) \leq M$ for some $M \geq 0$ and for any distributions p and q , the difference between the expected value of h under the distributions p and q is bounded by a function of the KL-divergence between these distributions as follows:

$$\begin{aligned}\mathbb{E}_q(h(I_T)) - \mathbb{E}_p(h(I_T)) &\leq \int_{q(I_T) > p(I_T)} h(I_T)(q(I_T) - p(I_T)) dI_T \\ &\leq M \int_{q(I_T) > p(I_T)} q(I_T) - p(I_T) dI_T \\ &= M \frac{1}{2} \int |q(I_T) - p(I_T)| dI_T \\ &\leq M \sqrt{\text{KL}(q||p)/2}.\end{aligned}\tag{B.4}$$

where $\text{KL}(q||p) = \int q(I_T) \log(q(I_T)/p(I_T)) dI_T$ is the KL-divergence between q and p and the last inequality is due to Pinsker's inequality [78], *i.e.*, $V(q, p) \leq \sqrt{\text{KL}(q||p)/2}$ where $V(q, p) = \int |q(I_T) - p(I_T)| dI_T / 2$ is the variational distance between q and p . The bound given in (B.4) is inspired by a similar bound obtained by [8] in the proof of Lemma A.1 for the case of discrete distribution in the context of non-stochastic multi-armed bandit problem.

Now, since $\sum_{t=1}^T \mathbb{1}\{x_t^\mu \geq \lambda_t\} \leq T$ and $\sum_{t=1}^T \mathbb{1}\{x_t^\mu < \lambda_t\} \leq T$, we use (B.4) to obtain

$$\mathcal{R}_T^\mu(f_1) \geq \frac{\epsilon}{2} \mathbb{E}_0 \left(\sum_{t=1}^T \mathbb{1}\{x_t^\mu \geq \lambda_t\} \right) - \frac{\epsilon}{2} T \sqrt{KL(f_0||f_1)/2},$$

and

$$\mathcal{R}_T^\mu(f_2) \geq \frac{\epsilon}{2} \mathbb{E}_0 \left(\sum_{t=1}^T \mathbb{1}\{x_t^\mu < \lambda_t\} \right) - \frac{\epsilon}{2} T \sqrt{KL(f_0||f_2)/2}.$$

Consequently,

$$\begin{aligned} \max_{i \in \{1,2\}} \mathcal{R}_T^\mu(f_i) &\geq \frac{1}{2} (\mathcal{R}_T^\mu(f_1) + \mathcal{R}_T^\mu(f_2)) \\ &\geq \frac{\epsilon}{4} \left(T - T \sqrt{KL(f_0||f_1)/2} - T \sqrt{KL(f_0||f_2)/2} \right). \end{aligned} \quad (\text{B.5})$$

For any $i \in \{0, 1, 2\}$, we can express the distribution of observations in terms of conditional distributions as follows;

$$\begin{aligned} f_i(I_T) &= \prod_{t=1}^T f_i(\pi_t, \lambda_t | x_t^\mu, I_{t-1}) f_i(x_t^\mu | I_{t-1}) \\ &= \prod_{t=1}^T f_i(\pi_t) f_\lambda(\lambda_t) f(x_t^\mu | I_{t-1}), \end{aligned}$$

where the second equality is due to the independence of λ_t and π_t from the past observations I_{t-1} , the bid x_t^μ , and from each other. Also, the distribution of x_t^μ given I_{t-1} does not depend on i . Consequently, for $i \in \{1, 2\}$,

$$\begin{aligned} KL(f_0||f_i) &= \int f_0(I_T) \log \left(\prod_{t=1}^T \frac{f_0(\pi_t)}{f_i(\pi_t)} \right) dI_T \\ &= \sum_{t=1}^T \int f_0(I_T) \log \left(\frac{f_0(\pi_t)}{f_i(\pi_t)} \right) dI_T \\ &= \sum_{t=1}^T \left(\frac{1}{2} \log \left(\frac{1/2}{1/2 + \epsilon} \right) + \frac{1}{2} \log \left(\frac{1/2}{1/2 - \epsilon} \right) \right) \\ &= -(T/2) \log(1 - 4\epsilon^2). \end{aligned}$$

Then, by (B.5) and by setting $\epsilon = T^{-1/2}/2\sqrt{5}$, we get

$$\begin{aligned} \max_{i \in \{1,2\}} \mathcal{R}_T^\mu(f_i) &\geq \frac{\epsilon T}{4} \left(1 - \sqrt{-T \log(1 - 4\epsilon^2)} \right) \\ &= \frac{\sqrt{T}}{8\sqrt{5}} \left(1 - \sqrt{-T \log(1 - 1/(5T))} \right) \\ &\geq \frac{\sqrt{T}}{16\sqrt{5}} \end{aligned}$$

where the last inequality follows from the fact that $-\log(1 - x) \leq (5/4)x$ for $0 \leq x \leq 1/5$.

Observe that the magnitude of the derivative of $r^{(0)}(x)$ is equal to $|\bar{\pi} - x|/\epsilon$ for $(1 - \epsilon)/2 \leq x \leq (1 + \epsilon)/2$ and 0 otherwise. So, for distributions f_1 and f_2 , $r^{(0)}(x)$ is Lipschitz continuous with Lipschitz constant $L = 3/2$ because $|\bar{\pi} - x|/\epsilon \leq 3/2$ for $(1 - \epsilon)/2 \leq x \leq (1 + \epsilon)/2$. Hence, assumptions (A1), (A2), and (A3) are satisfied for both distributions. \square

BIBLIOGRAPHY

- [1] R. Agrawal. The continuum-armed bandit problem. *SIAM J. Control Optim.*, 33(6):1926–1951, Nov. 1995.
- [2] Kareem Amin, Michael Kearns, Peter Key, and Anton Schwaighofer. Budget optimization for sponsored search: Censored learning in mdps. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI’12, pages 54–63, Arlington, Virginia, United States, 2012. AUAI Press.
- [3] Kareem Amin, Afshin Rostamizadeh, and Umar Syed. Learning prices for repeated auctions with strategic buyers. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1169–1177. Curran Associates, Inc., 2013.
- [4] T. W. Anderson and J. B. Taylor. Some experimental results on the statistical properties of least squares estimates in control problems. *Econometrica*, 44(6):1289–1302, Nov. 1976.
- [5] C. Antal, O. Granichin, and S. Levi. Adaptive autonomous soaring of multiple uavs using simultaneous perturbation stochastic approximation. In *49th IEEE Conf. Decision and Control (CDC)*, pages 3656–3661, 2010.
- [6] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.
- [7] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 322–331, Oct 1995.
- [8] Peter Auer, Nicol Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002.
- [9] Y. Aviv and A. Pazgal. A partially observed markov decision process for dynamic pricing. *Management Sci.*, 51(9):1400–1416, Sep. 2005.
- [10] M. Sevi Baltaoglu, Lang Tong, and Qing Zhao. Online learning of optimal bidding strategy in repeated multi-commodity auctions. In I. Guyon, U. V.

- Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4507–4517. Curran Associates, Inc., 2017.
- [11] S. Baltaoglu, L. Tong, and Q. Zhao. Online learning and optimization of markov jump linear models. In *41st IEEE Int. Conf. Acoust., Speech, and Signal Process.*, 2016.
 - [12] S. Baltaoglu, L. Tong, and Q. Zhao. Online learning and pricing for demand response in smart distribution networks. In *2016 IEEE Statistical Signal Processing Workshop (SSP)*, pages 1–5, June 2016.
 - [13] Sevi Baltaoglu, Lang Tong, and Qing Zhao. Algorithmic bidding for virtual trading in electricity markets.
 - [14] Sevi Baltaoglu, Lang Tong, and Qing Zhao. Online learning and optimization of markov jump affine models.
 - [15] R. J. Balvers and T. F. Cosimano. Actively learning about demand and the dynamics of price adjustment. *Econ. J.*, 100(402):882–898, Sep. 1990.
 - [16] D. Bertsimas and G. Perakis. Dynamic pricing: A learning approach. In *Mathematical and Computational Models for Congestion Charging*, pages 45–79. Springer, Boston, MA, USA, 2006.
 - [17] O. Besbes, Y. Gur, and A. Zeevi. Non-stationary stochastic optimization. *Oper. Res.*, 63(5):1227–1244, Sep.-Oct. 2015.
 - [18] John Birge, Ali Hortacsu, Ignacia Mercadal, and Michael Pavlin. Limits to arbitrage in electricity markets: A case study of miso. *MIT Center for Energy and Environmental Policy Research*, 2017.
 - [19] B. Bobrovsky and M. Zakai. A lower bound on the estimation error for markov processes. *IEEE Trans. Autom. Control*, 20(6):785–788, Dec. 1975.
 - [20] S. Borenstein, M. Jaske, and A. Rosenfeld. Dynamic pricing, advanced metering, and demand response in electricity markets. Technical report, 2002.
 - [21] Severin Borenstein, James Bushnell, Christopher R. Knittel, and Catherine Wolfram. Inefficiencies and market power in financial arbitrage: A study of california’s electricity markets*. *J. Ind. Econ.*, 56(2):347–378, June 2008.

- [22] J. Broder and P. Rusmevichientong. Dynamic pricing under a general parametric choice model. *Oper. Res.*, 60(4):965–980, Jul.-Aug. 2012.
- [23] Sbastien Bubeck and Nicol Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.
- [24] Nicolò Cesa-Bianchi, Yoav Freund, David P. Helmbold, David Haussler, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. In *Proceedings of the Twenty-fifth Annual ACM Symposium on Theory of Computing*, STOC '93, pages 382–391, New York, NY, USA, 1993. ACM.
- [25] Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.
- [26] E. W. Cope. Regret and convergence bounds for a class of continuum-armed bandit problems. *IEEE Trans. Autom. Control*, 54(6):1243–1253, Jun. 2009.
- [27] O.L.V. Costa, M.D. Fragoso, and R.P. Marques. *Discrete-Time Markov Jump Linear Systems*. Probability and Its Applications. Springer, 2005.
- [28] Constantinos Daskalakis and Vasilis Syrgkanis. Learning in auctions: Regret is hard, envy is easy. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 219–228, 2016.
- [29] A. V. den Boer and B. Zwart. Simultaneously learning and optimizing using controlled variance pricing. *Management Sci.*, 60(3):770–783, Mar. 2014.
- [30] A. Doucet, N. J. Gordon, and V. Krishnamurthy. Particle filters for state estimation of jump markov linear systems. *IEEE Trans. Signal Process.*, 49(3):613–624, Mar 2001.
- [31] Miroslav Dudik, Nika Haghtalab, Haipeng Luo, Robert E. Shapire, Vasilis Syrgkanis, and Jennifer Wortman Vaughan. Oracle-efficient online learning and auction design. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 528–539, 2017.
- [32] Krzysztof Dudziski and Stanisaw Walukiewicz. Exact methods for the knapsack problem and its generalizations. *Eur. J. Oper. Res.*, 28(1):3 – 21, 1987.

- [33] Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: Gradient descent without a gradient. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '05*, pages 385–394, Philadelphia, PA, USA, 2005. Society for Industrial and Applied Mathematics.
- [34] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceedings of the Second European Conference on Computational Learning Theory, EuroCOLT '95*, pages 23–37, London, UK, UK, 1995. Springer-Verlag.
- [35] R. D. Gill and B. Y. Levit. Applications of the van trees inequality: A bayesian cramer-rao bound. *Bernoulli*, 1(1/2):59–79, Mar.-Jun. 1995.
- [36] T. Guler, G. Gross, E. Litvinov, and R. Coutu. On the economics of power system security in multi-settlement electricity markets. *IEEE Trans. Power Syst.*, 25(1):284–295, Feb. 2010.
- [37] William W. Hogan. Virtual bidding and electricity market design. *Electricity J.*, 29(5):33 – 47, June 2016.
- [38] Akshaya Jha and Frank A. Wolak. Testing for market efficiency with transactions costs: An application to convergence bidding in wholesale electricity markets, 2015.
- [39] L. Jia, Q. Zhao, and L. Tong. Retail pricing for stochastic demand with unknown parameters: An online machine learning approach. In *2013 51st Annu. Allerton Conf. Communication, Control, and Computing (Allerton)*, pages 1353–1358, 2013.
- [40] Hans Kellerer, Ulrich Pferschy, and David Pisinger. The multiple-choice knapsack problem. In *Knapsack Problems*, chapter 11, pages 317–347. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.
- [41] N. B. Keskin and A. Zeevi. Chasing demand: Learning and earning in a changing environment. Working paper. Available at <http://ssrn.com/abstract=2389750>, 2013.
- [42] N. B. Keskin and A. Zeevi. Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Oper. Res.*, 62(5):1142–1167, Sep.-Oct. 2014.

- [43] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.*, 23(3):462–466, Sep. 1952.
- [44] R. Kleinberg and T. Leighton. The value of knowing a demand curve: bounds on regret for online posted-price auctions. In *Proc. 44th Annu. IEEE Symp. Foundations Computer Science*, pages 594–605, 2003.
- [45] R. D. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems 17*, pages 697–704, 2005.
- [46] Robert Kleinberg, Alexandru Niculescu-Mizil, and Yogeshwer Sharma. Regret bounds for sleeping experts and bandits. In *21st Conference on Learning Theory*, pages 425–436, 2008.
- [47] Robert Kleinberg and Aleksandrs Slivkins. Sharp dichotomies for regret minimization in metric spaces. In *Proc. 21th Annu. ACM-SIAM Symp. Discrete Algorithms*, pages 827–846, 2010.
- [48] Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Bandits and experts in metric spaces. *CoRR*, abs/1312.1277, 2013.
- [49] Robert D. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 697–704. MIT Press, 2005.
- [50] Walid Krichene, Maximilian Balandat, Claire Tomlin, and Alexandre Bayen. The hedge algorithm on a continuum. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37, ICML’15*, pages 824–832. JMLR.org, 2015.
- [51] T. L. Lai. Asymptotically efficient adaptive control in stochastic regression models. *Adv. Appl. Math.*, 7(1):23 – 45, Mar. 1986.
- [52] T. L. Lai and H. Robbins. Adaptive design and stochastic approximation. *Ann. Statist.*, 7(6):1196–1221, Nov. 1979.
- [53] T. L. Lai and C. Z. Wei. Asymptotically efficient self-tuning regulators. *SIAM J. Control Optim.*, 25(2):466–481, Mar. 1987.
- [54] T.L. Lai and H. Robbins. Iterated least squares in multiperiod control. *Adv. Appl. Math.*, 3(1):50 – 73, Mar. 1982.

- [55] T.L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4 – 22, 1985.
- [56] Ruoyang Li, Alva J. Svoboda, and Shmuel S. Oren. Efficiency impact of convergence bidding in the california electricity market. *J. Regul. Econ.*, 48(3):245–284, Dec. 2015.
- [57] Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212 – 261, 1994.
- [58] M. Lobo and S. Boyd. Pricing and learning with uncertain demand. presented at the INFORMS Revenue Management Conf., 2003.
- [59] A. Logothetis and V. Krishnamurthy. Expectation maximization algorithms for map estimation of jump markov linear systems. *IEEE Trans. Signal Process.*, 47(8):2139–2156, Aug 1999.
- [60] Harry Markowitz. Portfolio selection. *J. Finance*, 7(1):77–91, Mar. 1952.
- [61] Jonathan Mather, Eilyan Bitar, and Kameshwar Poolla. Virtual bidding: Equilibrium, learning, and the wisdom of crowds. *IFAC-PapersOnLine*, 50(1):225 – 232, 2017. 20th IFAC World Congress.
- [62] Colin McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics, 1989: Invited Papers at the Twelfth British Combinatorial Conference*, pages 148–188. Cambridge University Press, Cambridge, 1989.
- [63] Paul Milgrom. *Putting auction theory to work*. Cambridge University Press, 2004.
- [64] Mehryar Mohri and Andres Munoz. Optimal regret minimization in posted-price auctions with strategic buyers. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1871–1879. Curran Associates, Inc., 2014.
- [65] U.S. Department of Energy. Benefits of demand response in electricity markets and recommendations for achieving them. Technical report, 2006.
- [66] John E. Parsons, Cathleen Colbert, Jeremy Larrieu, Taylor Martin, and Erin Mastrangelo. Financial arbitrage and efficient dispatch in wholesale elec-

- tricity markets. *MIT Center for Energy and Environmental Policy Research No. 15-002*, 2015.
- [67] David B. Patton, Pallas LeeVanSchaick, and Jie Chen. 2014 state of the market report for the new york iso markets. Technical report, May 2015.
 - [68] PJM. Virtual transactions in the pjm energy markets, 2015.
 - [69] H. Robbins and D. Siegmund. *A Convergence Theorem for Non Negative Almost Supermartingales and Some Applications*, pages 111–135. Springer, New York, NY, USA, 1985.
 - [70] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407, 09 1951.
 - [71] P. Rusmevichientong and J. N. Tsitsiklis. Linearly parameterized bandits. *Math. Oper. Res.*, 35(2):395–411, May 2010.
 - [72] Celeste Saravia. Speculative trading and market performance: The effect of arbitrageurs on efficiency and market power in the new york electricity market. *Center for the Study of Energy Markets*, 2003.
 - [73] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.
 - [74] J. C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Autom. Control*, 37(3):332–341, Mar. 1992.
 - [75] Wenyuan Tang, Ram Rajagopal, Kameshwar Poolla, and Pravin Varaiya. Impact of virtual bidding on financial and economic efficiency of wholesale electricity markets. Working paper.
 - [76] Wenyuan Tang, Ram Rajagopal, Kameshwar Poolla, and Pravin Varaiya. Model and data analysis of two-settlement electricity market with virtual bidding. In *2016 IEEE 55th Conf. Decision and Control*, pages 6645–6650, 2016.
 - [77] Long Tran-Thanh, Lampros Stavrogiannis, Victor Naroditskiy, Valentin Robu, Nicholas R Jennings, and Peter Key. Efficient regret bounds for online bid optimisation in budget-limited sponsored search auctions. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*,

UAI'14, pages 809–818, Arlington, Virginia, United States, 2014. AUA Press.

- [78] Alexandre B. Tsybakov. Lower bounds on the minimax risk. In *Introduction to Nonparametric Estimation*, chapter 2, pages 77–135. Springer New York, New York, NY, 2009.
- [79] Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances Neural Inform. Process. Syst.* 4, pages 831–838. 1992.
- [80] Jonathan Weed, Vianney Perchet, and Philippe Rigollet. Online learning in repeated auctions. In *29th Annu. Conf. Learning Theory*, pages 1562–1583, 2016.
- [81] C.K. Woo, J. Zarnikau, E. Cutter, S.T. Ho, and H.Y. Leung. Virtual bidding, wind generation and california's day-ahead electricity forward premium. *Electricity J.*, 28(1):29 – 48, Feb. 2015.
- [82] G. Yin, C. Ion, and V. Krishnamurthy. How does a stochastic optimization/approximation algorithm adapt to a randomly evolving optimum/root with jump markov sample paths. *Math. Program.*, 120(1):67–99, Aug. 2009.
- [83] Q. Zhao, L. Tong, A. Swami, and Y. Chen. Decentralized cognitive mac for opportunistic spectrum access in ad hoc networks: A pomdp framework. *IEEE Journal on Selected Areas in Communications*, 25(3):589–600, April 2007.
- [84] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pages 928–935. AAAI Press, 2003.